UNIVERSITÀ DEGLI STUDI DI PADOVA

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

UNIVERSITÀ DEGLI STUDI DI PADOVA

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

CORSO DI LAUREA MAGISTRALE IN BIOINGEGNERIA

"Detection of Epileptic Seizures from Audio Recordings"

Relatrice:
Maria RUBEGA, PhD

Correlatore:
Dr. Mario CHAVEZ (CNRS, France)

Laureanda:
Isabella IOVIENO

Correlatore:
Edoardo PASSAROTTO, PhD

ANNO ACCADEMICO 2023 – 2024

23/04/24

ABSTRACT

The health risks associated with epileptic seizures make early detection and effective intervention essential. Seizures detection devices are a useful tool for epilepsy monitoring. Accurate detectors can alert caregivers and reduce the distress of individuals experiencing nocturnal seizures. For this reason, there is an urgent need for non-invasive, accurate, affordable, and easily wearable technologies that can be used both in clinical and home monitoring settings. Thanks to recent advancements in machine learning and audio technologies, audio recording systems are considered a valuable option for remote monitoring.

This research project presents a preliminary study on detection of epileptic seizures from audio recordings acquired in the patients' room at the Epilepsy Unit of La Pitié-Salpêtrière hospital (Paris, France). Here, we analysed more than seven hours of audio data extracted during both seizure (ictal) and seizure-free (interictal) epochs of 5 hospitalized patients. Every audio-video recording was hand-labelled, to create a carefully selected dataset. The manual labelling process has enhanced the availability of high-quality datasets, a vital resource that is frequently lacking, benefiting not only the study's immediate objectives, but also the larger epilepsy research community.

After the video recordings have been carefully labelled, the audio data were segmented in overlapped 10-s windows and pre-processed to extract relevant features. The proposed approach integrated standard features from: i) the time-frequency domain (on Mel's scale) such as the spectral centroid, spectral entropy, spectral roll-off point, the cepstral coefficients; ii) the time domain, which include, for example, zero-crossing rate and energy. The extracted features served as the foundation for the application of different statistical models. Firstly, I used the Mahalanobis distance to measure the statistical dissimilarity between features vector in a current audio segment from those obtained during some seizure-free periods. By considering possible feature correlations, the Mahalanobis distance offered a reliable measure of dissimilarity that is essential to detect significant deviations from the reference class (here the seizure-free epochs). In a second part, we applied different machine learning models (decision trees, SVMs, k-NN and neural networks), which were trained on the manually labelled dataset and evaluated with a k-fold cross-validation to distinguish ictal and interictal segments.
Performance metrics such as accuracy, precision, recall and F1-score (a better adapted metric for imbalanced datasets) resulting from a k-fold cross-validation were used to assess the performance of the detector. To address the class imbalance of the dataset, we also applied a statistical technique (SMOTE) to over-sample the minority class. Results showed that class

imbalance yield overoptimistic results in general. When the imbalanced was corrected, results indicated that some methods (Mahalanobis-based detector, k-NN and decision trees) performed well and allowed to accurately identify the epoch with seizures.

Our results suggest that the proposed method for detecting seizures offers a promising starting point for further investigation and refinement. The study not only contributes to the development of reliable seizure detection systems but also addresses critical considerations for practical implementation. Beyond the algorithmic framework, the research offers insights into the potential applications of audio-based detection in real-time seizure monitoring, presenting a non-intrusive and cost-effective alternative for long-term patient care. If these preliminary results will be confirmed in larger datasets, the proposed approach may be a useful contribution to healthcare technology, especially in neurological research.

Keywords: epilepsy, automatic seizure detection, audio recordings, machine learning

SOMMARIO

I rischi per la salute associati alle crisi epilettiche rendono essenziale la loro individuazione precoce e un intervento efficace. I dispositivi di rilevamento delle crisi sono uno strumento utile per il monitoraggio dell'epilessia. Dei buoni rilevatori possono allertare gli operatori sanitari o i caregiver e ridurre il disagio degli individui che soffrono di convulsioni notturne. Per questo motivo, c'è urgente bisogno di tecnologie non invasive, accurate, accessibili e facilmente indossabili che possano essere impiegate sia in ambienti clinici che domestici. Grazie ai recenti progressi nelle tecnologie di apprendimento automatico e nelle tecnologie di analisi del suono, i sistemi di monitoraggio audio sono pensati come una valida opzione per il monitoraggio dei pazienti epilettici.

Questo progetto di ricerca presenta uno studio preliminare sulla rilevazione delle crisi epilettiche da registrazioni audio acquisite nelle stanze dei pazienti presso l'Unità di Epilessia dell'ospedale La Pitié-Salpêtrière (Parigi, Francia). A tal fine, abbiamo analizzato più di sette ore di dati audio estratti durante periodi sia di crisi epilettiche (ictale) che senza crisi (interittale) di 5 pazienti ospedalizzati. Ogni registrazione audio-video è stata etichettata manualmente, evidenziando lo sforzo per creare un dataset accuratamente selezionato. Il processo manuale di etichettatura permette di incrementare la disponibilità di dataset di alta qualità, una risorsa vitale spesso insufficiente, dalla quale possono beneficiare non solo gli obiettivi immediati dello studio ma anche la comunità di ricerca sull'epilessia.

Dopo che le registrazioni video sono state accuratamente etichettate, i dati audio vengono segmentati in finestre sovrapposte di 10 secondi e pre-elaborati per estrarre caratteristiche rilevanti.
L'approccio proposto integra caratteristiche standard da: i) il dominio tempo-frequenza (su scala di Mel) come il centroide spettrale, l'entropia spettrale, il punto di roll-off spettrale, i coefficienti cepstrali; ii) il dominio temporale, che include il tasso di attraversamento dello zero e l'energia media quadratica. Le caratteristiche estratte fungono da base per l'applicazione di diversi modelli statistici. Prima di tutto, ho utilizzato la distanza di Mahalanobis per misurare la dissimilarità statistica tra vettori di caratteristiche in un segmento audio corrente e quelli ottenuti durante alcuni periodi senza crisi. Considerando possibili correlazioni tra le caratteristiche, la distanza di Mahalanobis offre una misura affidabile di dissimilarità essenziale per rilevare deviazioni significative dalla classe di riferimento (qui epoche senza crisi). In una seconda parte, abbiamo applicato diversi modelli di apprendimento automatico (alberi decisionali, SVM, kNN e reti neurali), che sono stati addestrati sul dataset etichettato

manualmente e valutati con una cross-validazione k-fold per distinguere tra segmenti ictali e interittali.

Metriche di prestazione come accuratezza, precisione, richiamo e misura F1 (una metrica più adatta per dataset sbilanciati) ottenute da una cross-validazione k-fold sono state utilizzate per valutare le prestazioni del rilevatore. Per affrontare lo sbilanciamento delle classi nel dataset, abbiamo applicato anche una tecnica statistica (SMOTE) per sovracampionare la classe minoritaria.

I risultati mostrano che lo sbilanciamento delle classi produce generalmente risultati troppo ottimistici. Quando lo sbilanciamento viene corretto, i risultati indicano che alcuni metodi (rilevatore basato su Mahalanobis, k-NN e alberi decisionali) hanno buone prestazioni e consentono di identificare accuratamente l'epoca con le crisi.

I nostri risultati suggeriscono che il metodo proposto per la rilevazione delle crisi costituisce un punto di partenza promettente per ulteriori indagini e affinamenti. Lo studio contribuisce non solo allo sviluppo di sistemi affidabili di rilevamento delle crisi ma affronta anche considerazioni critiche per l'implementazione pratica. Oltre al quadro algoritmico, la ricerca offre spunti sulle potenziali applicazioni della rilevazione basata sull'audio nel monitoraggio in tempo reale delle crisi, presentando un'alternativa non invasiva ed economica per l'assistenza a lungo termine ai pazienti. Se i risultati preliminari saranno confermati su set di dati più ampi, l'approccio proposto potrebbe rappresentare un contributo utile alla tecnologia sanitaria, in particolare nella ricerca neurologica.

Parole chiave: epilessia, rilevamento automatico delle crisi epilettiche, registrazioni audio, apprendimento automatico

ACKNOWLEDGEMENTS

I extend my heartfelt gratitude to Mario Chavez and all members of the team for granting me the invaluable opportunity to work at the Institute du Cerveau in Paris. This experience has been profoundly enriching, and I will forever treasure it.

I am also thankful to my advisor, Maria Rubega, and my co-advisor, Edoardo Passarotto. Your consistent guidance and motivation have been instrumental throughout my journey, and I deeply appreciate it.

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION 1

1.1 The epilepsy 1

1.2 Detection 5

1.3 Seizure Detection Methods 6

1.4 Rationale for Audio-Based Seizure Detection 10

1.5 State of the art: audio-based seizure detection 11

1.5.1 Feature extraction 12

1.5.2 Machine learning 13

1.6 Objective of the study 15

CHAPTER 2: MATERIALS AND METHODS 19

2.1 A word about the hosting laboratory 19

2.2 Data Collection 20

2.2.1 Instrumentation 20

2.2.2 Video Selection 21

2.3 Audio extraction from video recordings 23

2.4 Exploratory Data Analysis 23

2.5 Data Pre-Processing 27

2.5.1 Audio segmentation and windowing 28

2.5.2 Labelling 29

2.6 Features Extraction 30

2.7 Dataset Characteristics 38

2.8 Statistical Analysis 39

2.8.1 Mahalanobis distance 39

2.8.2 Machine Learning 40

2.8.2.1 k-Nearest Neighbours 40

2.8.2.1.1 Weighted k-Nearest Neighbours 41

2.8.2.2 Linear Support Vector Machine 42
2.8.2.3 Decision Trees 44
2.8.2.4 Neural Networks 46
2.9 Training and testing procedures 49
2.10 Performance Evaluation 50
2.11 Effect of imbalanced data 52
2.11.1 Synthetic Minority Over-Sampling Technique (SMOTE) 53

CHAPTER 3: EXPERIMENTAL RESULTS 56
3.1 Features Visualization 56
3.2 Results of Mahalanobis Distance 57
3.3 Performance Metrics on the original imbalanced data 58
3.3.1 Performance metrics for Patient 3132 (imbalanced data) 58
3.3.2 Performance metrics averaged among all patients (imbalanced data) 64
3.4 Performance Metrics on balanced data (SMOTE algorithm) 65
3.4.1 Performance metrics for Patient 3132 (balanced data) 66
3.4.2 Performance metrics averaged among all patients (balanced data) 67

CHAPTER 4: DISCUSSION 70
4.1 Prior to data augmentation 70
4.2 Following data augmentation to balance the classes 73
4.3 Limitations 75

CHAPTER 5: CONCLUSIONS 78
5.1 Future Developments 78
BIBLIOGRAPHY 81

LIST OF FIGURES

Figure 1. Illustration of focal, generalized and unknown onset, with percentage of occurrence [7]. 2
Figure 2. Illustration of epilepsy treatments options currently available [13]. 4
Figure 3. Pipeline of the proposed approach for the detection of epileptic seizures from audio recordings. From the video recordings, we extracted the corresponding audio recordings, then we performed pre-processing steps such as visual investigation, audio segmentation and windowing and labelling of windows. We extracted features from each segment, performed the detection task employing machine learning models and finally, we validated the models..... 16
Figure 4. Institut du Cerveau (ICM – Paris Brain Institute)..... 19
Figure 5. a) AXIS M5525–E PTZ Network Camera with support to install it on the wall of the patient's room and b) AXIS T8351 Mk II Microphone 3.5 mm. 21
Figure 6. A patient in their hospital room at Pitié-Salpêtrière, during the daytime, undergoing a VEEG procedure. 22
Figure 7. A patient in their hospital room at Pitié-Salpêtrière, during the nighttime, undergoing a VEEG procedure. 23
Figure 8. Plot of waveform, spectrogram and Mel spectrogram of video 261 of patient 3132 with seizure epoch markers. Seizure occurs between the second 56 and the second 80. 24
Figure 9. Visual representation of pre-processing steps, features extraction step, creation of the dataset step and detection step. 28
Figure 10. Visualization of Mel-frequency cepstral coefficients (MFCCs) of video 261 of patient 3132. Ictal period occurs from 56 sec to 80 sec. 36
Figure 11. Visual result of employing the sum of Mel-frequency cepstral coefficients (MFCCs) of video 261 of patient 3132, over time. Ictal period occurs from 56 sec to 80 sec. 36
Figure 12. Plot of waveform of audio signal from video 261 of patient 3132 and some of the features extracted from it, such as: mfccs, spectral entropy, spectral rolloff point, zero crossing rate and root mean square energy. In red the markers for the beginning and the end of the seizure period. 37
Figure 13. Visual representation of how the algorithm of k-Nearest-Neighbours works. 40
Figure 14. Visual representation of how the algorithm of linear support vector machine model works. 44
Figure 15. Visual representation of a decision tree model. 45
Figure 16. Visual representation of a 1-hidden layer neural network model. 48
Figure 17. Visual representation of a 2-hidden layers neural network model..... 48

Figure 18. Visualisation of Synthetic Minority Over-Sampling Technique (SMOTE), a) imbalanced dataset priore to data augmentation and b) dataset after SMOTE. Blue circles represent non-seizure data, red circles seizure data and green circles new synthetic seizure data. 53
Figure 19. Boxplot of time-domain features (mean, zero crossing rate and root mean square energy) obtained from audio recordings of patient 3132 divided among its different sound labels: seizure (ictal epoch), silence (minimal sound activity) and other (undefined sounds). 56
Figure 20. Boxplot of feature zero crossing rate obtained from all patients and distributed among all sound labels present in all audio recordings. 57
Figure 21. Mahalanobis distance of audio recordings extracted from videos 260 and 261 of patient 3132 with seizure start and end markers in red. 57
Figure 22. Local density of datapoints of patient 3132 obtained with k-NN model (k=5) in red the seizure start and end markers. 58
Figure 23. Confusion matrix of k-NN model generated from the test set (15%) of patient 3132. The model utilized time-frequency domain features as input, where each signal window yielded a feature vector. The size of the feature vector depended on the number of frames (partitions) and the degree of overlap used to estimate the Mel spectrogram and the related parameters..59
Figure 24. Scatter plot of predictions of model Weighted k-Nearest Neighbours implemented with time-domain features of patient 3132 as input. Blue points indicate observations from the "non-seizure" class, while red points indicate the "seizure" class. 60
Figure 25. Confusion matrices of weighted k-NN, linear SVM, one-layer NN, bi-layer NN and decision tree models generated from the test set of patient 3132 utilizing time-frequency domain features as input (original imbalanced dataset). 62
Figure 26. Confusion matrices of weighted k-NN, linear SVM, one-layer NN, bi-layer NN and decision tree models generated from the test set of patient 3132 utilizing time-frequency domain features as input (balanced dataset). 66

LIST OF TABLES

Table 1. ID of patients, time of the day in which the seizures start, day period in which the seizures occur, state of vigilance of the patients, and seizure types, and overall duration of the audio recordings. 21

Table 2. Parameters to calculate both the spectrogram and Mel spectrogram. Window length refers here to the segment's length involved in the STFT algorithm. 27

Table 3. Table of patient 3152 containing information about the time of interest of each sound event appearing in each video, its corresponding label, the name of the videos and the folder where the videos are. 30

Table 4. Quantified seizure and non-seizure periods for each patient in terms of both duration (seconds) and the number of windows. 38

Table 5. Accuracy, Precision, Recall and F1 Score obtained for k-NN model and patient 3132 both using time-domain features and time-frequency domain features (original imbalanced dataset). 59

Table 6. Accuracy obtained after training and test of models weighted k-NN, linear SVM, decision tree, 1-Layer NN and 2-Layers NN performed with time-domain features and time-frequency features as input for patient 3132 (original imbalanced dataset). 63

Table 7. Precision, Recall, F1 Score and Area Under the Curve (AUC) obtained after test of models weighted k-NN, linear SVM, decision tree, 1-Layer NN and 2-Layers NN performed with time-domain features and time-frequency features as input (original imbalanced dataset). 64

Table 8. Mean values of Accuracy, Precision, Recall, F1 Score and AUC averaged among all patients with time-domain features and time-frequency domain features as input of weighted k-NN, linear SVM, decision tree, 1-layer NN and 2-layers NN (original imbalanced dataset).. 65

Table 9. Minimum and maximum values of Accuracy, Precision, Recall, F1 Score and AUC registered among all patients with time-domain features and time-frequency domain features as input of weighted k-NN, linear SVM, decision tree, 1-layer NN and 2-layers NN (original imbalanced dataset). 65

Table 10. Values of Accuracy, Precision, Recall, F1 Score and AUC obtained for patient 3132 with time-domain features and time-frequency domain features as input of k-NN, weighted k-NN, linear SVM, decision tree, 1-layer NN and 2-layers NN, after balancing the dataset implementing the SMOTE algorithm. 67

Table 11. Values of Accuracy, Precision, Recall, F1 Score and AUC obtained averaged among all patients with time-domain features and time-frequency domain features as input of k-NN,

weighted k-NN, linear SVM, decision tree, 1-layer NN and 2-layers NN, after balancing the dataset implementing the SMOTE algorithm. 68
Table 12. Minimum and maximum values of Accuracy, Precision, Recall, F1 Score and AUC registered among all patients with time-domain features and time-frequency domain features as input of weighted k-NN, linear SVM, decision tree, 1-layer NN and 2-layers NN (balanced dataset). 68

CHAPTER 1: INTRODUCTION

1.1 The epilepsy

Epilepsy is a neurological disorder characterized by recurrent seizures, which result from brief episodes of abnormally high, irregular, sparse, involuntary, and hyper-synchronized brain neuronal activity [1]. Epileptic seizures arise from an abnormal hyperexcitability of neurons, which generate and propagate abnormal and excessive electrical activity through the cortex.

Accordingly to the World Health Organization, epilepsy impacts nearly 60 million individuals (approximately 1% of the global population) and witnessing 2.5 million new cases annually [2]. Individuals with epilepsy face an elevated mortality risk, ranging two to three times higher than the general population, resulting in a mean annual mortality rate of 1% [3].

Epilepsy affects people of all ages, genders, and backgrounds, making it one of the most common neurological conditions globally. The underlying causes of epilepsy can be diverse, including genetic factors, brain injuries, infections, or developmental disorders. In some cases, the cause may remain unknown [1].

Depending on how the seizures begin and spread inside the brain, a wide range of behaviours may be observed during epileptic seizures [4]. Seizures can manifest in various ways, ranging from momentary lapses in awareness to full-body convulsions.

Seizures are classified into two types: unprovoked (spontaneous) and provoked. Unprovoked seizures occur in cases of persistent brain disease, such as epilepsy [5]. On the other hand, provoked seizures occur when certain factors trigger them in an otherwise healthy brain, such as acute metabolic processes, acute neurologic insult, drugs, and excessive physiological conditions. Every patient has its own epileptic condition strongly affected by her/his clinical history, prognosis, and likely response to particular treatments [5].

The standard clinical classification of epileptic seizures considers three families: focal (affecting a partial section of the cortex), generalized (involving large areas of the cortex) and difficult to define seizures [6].

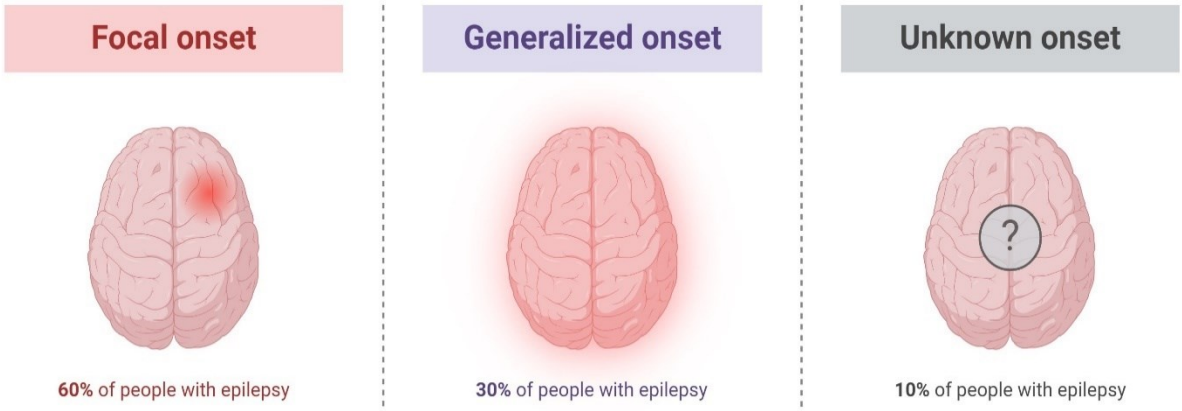

Figure 1. Illustration of focal, generalized and unknown onset, with percentage of occurrence [7].

It is important to have a preliminary distinction among focal, generalized, and unknown onset seizures since each type of seizure has a different set of symptoms and response to treatment [6]. In fact, the main criteria used to categorize seizures are the symptoms that are present during the seizure [6].

Focal or partial seizures mainly originate within networks limited to one hemisphere [8], as shown in **Figure 1**. Focal seizures present as either simple partial seizures or complex partial seizures. Simple partial seizures are localized in specific and small brain areas, often inducing unusual sensations or movements/twitching without a loss of awareness. In contrast, complex partial seizures impact larger brain regions, resulting in altered consciousness, repetitive behaviours, automatisms, epileptic spasms, hyperkinetic and myoclonic [9].

Generalized epileptic seizures are believed to start from a limited point of the brain before quickly propagate to the whole cortex [10]. These brain networks may involve cortical and subcortical structures. While the onset of a seizure may seem to be focused in one area, the location and lateralization may not be consistent from one seizure to the next. Generalized seizures can be divided into tonic-clonic seizures (*grand mal*) and absence seizures (*petit mal*). The first ones are characterized by convulsive movements, which are involuntary muscle contractions involving tonic (muscle stiffness) and clonic (rhythmic jerking) phases. They can cause cry-outs, loss of consciousness, falling to the ground and muscle spasms. The second ones start without convulsive movements and can cause rapid blinking or a few seconds of staring into space. Another subgroup is formed by the secondary generalized seizures, which begin in one limited part of the brain and then spread to both hemispheres.

Although most seizures last less than five minutes, they can sometimes be preceded by a prodromal phase and followed by a protracted postictal period in which the patient gradually returns to baseline [6].

Most seizures stop on their own without any intervention. In rare cases, seizures can develop into a life-threatening condition called sudden unexpected death in epilepsy (SUDEP) or status epilepticus (prolonged seizures). It is important to note that not all convulsions are of epileptic origin, and not all seizures cause convulsions. For example, a syncopal convulsion may be caused by a drop in blood pressure, not abnormal brain activity [5]. Numerous variables, including the type of seizure, the age at which seizures begin, family history, physical examination findings, ictal and interictal electroencephalography (EEG), and neurologic imaging, are used to characterize epileptic syndromes [5].

Epilepsy generally has two peaks of incidence during the course of life of patients: the first during childhood, the second in the elderly population, after the age of 65. Overall, complex partial seizures (focal) are the most common type of seizure across age groups. Overall, up to 61% of people with epilepsy experience focal seizures [11], 30% of people with epilepsy experience generalized seizures and 10% experience unknown onset seizures, as reported in **Figure 1**.

The main issue, regardless of clinical classification, is the unpredictability of seizures. In fact, seizures events are unpredictable and can happen at inconvenient, unpleasant, or even harmful times, except for a small number of cases with reflex epilepsy (i.e. syndrome in which seizures are triggered by some stimulus), in which case seizures are triggered by external stimuli, internal mental processes, or both [12]. Frustration is increased by the fact that patients typically have no control over when seizures start or stop [6]. The unpredictable nature of seizures can significantly impact an individual's daily life, influencing aspects such as education, employment, and overall quality of life. For many patients, this is the most upsetting part of living with epilepsy.

Though there is no cure for epilepsy at present, its symptoms can be managed through medication, lifestyle changes, Vagus nerve stimulation, deep brain stimulation, and in the case of focal seizures only, surgical interventions (**Figure 2**). Ongoing research into epilepsy is leading to a better understanding of the condition and improved treatment options for those living with it.

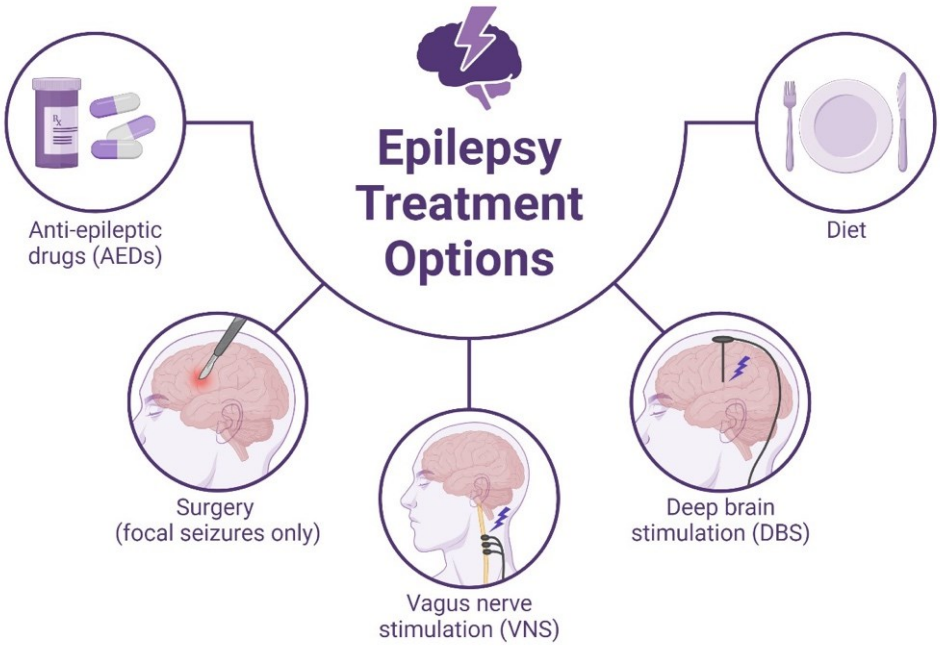

Figure 2. *Illustration of epilepsy treatments options currently available* [13].

Epilepsy is currently treated with various medications to prevent or reduce the frequency of seizures and improve the quality of life of patients. However, some people with epilepsy are resistant to anti-epileptic medication. Indeed, up to 20 to 40% of individuals with epilepsy have refractory epilepsy, meaning that standard treatments are ineffective in controlling their seizures [14]. When epileptic patients do not respond to drug treatment, surgery to remove or disrupt the part of the brain causing seizures (known as the epileptogenic zone or EZ) may be an option for achieving seizure freedom or reducing seizure frequency in focal epilepsy.
However, individuals with generalized epilepsy are often not considered candidates for epilepsy surgery [10]. Alarmingly, also surgery is ineffective for many patients [3] and a considerable number of people with epilepsy still experience seizures [15]. This means that they have to deal with seizures for an extended period of time, possibly for the rest of their lives [15].

Considering all that is mentioned above, at La Pitié-Salpêtrière Hospital patients who suffer from focal and focal to bilateral epilepsy and for whom both anti-epileptic medication did not work, are undergoing investigation utilizing intracranial electrodes. These procedures are planned to study brain activity during seizure episodes to try to better localize the neurons involved in seizure activity. Using this technique, scientists hope to learn more about the mechanisms underlying epileptic seizures and discover new treatment options.
Since seizure monitoring is essential for making therapeutic decisions, patients or caregivers are usually asked to keep records of seizure episodes through diaries. Many clinical trials in the

field of epilepsy use metrics generated from patient-reported seizure frequency as their primary outcome measure. Thus, reliable seizure documentation by patients or caregivers is crucial to the advancement of science and clinical practice. Studies have revealed, however, that patients' and caregivers' documentation of seizures is often inaccurate. The type of seizures being recorded influences how accurate the reporting of seizures is, and this might vary significantly within an individual over time [16]. There is therefore a need to explore other options to monitor patients affected by epilepsy. Additionally, and this concerns this research project, there is a need and growing interest to more effectively monitor and support epilepsy patients both at home and during their hospital stay.

1.2 Detection

From a statistical point of view, the problem of detection revolves around the task of identifying the presence or absence of specific events or phenomena within a given dataset. This covers a wide range of applications, from recognizing the occurrence of specific events to distinguishing instances of interest (or anomalies) from normal data. Detection using two classes involves the classification of instances into one of two categories based on specific criteria. In this context, a statistical model is trained to distinguish between two distinct classes, typically denoting the presence or absence of a particular event or condition.

In the context of epileptic seizure detection, the classes usually refer to different brain states: "seizure" and "non-seizure" or "ictal" and "inter-ictal" periods. The training process involves providing the model with representative samples of both classes, enabling it to learn discriminative features that differentiate between the two. Subsequently, during the detection phase, the model evaluates new instances and assigns them to one of the two classes based on the learned patterns. This approach is fundamental to a wide range of detection problems, providing a structured framework for decision-making in scenarios where outcomes can be categorized into two distinct classes.

Anomaly detection, also known as outlier detection, involves isolating a single class of data from other classes in the feature space [17]. It is a specialized branch of detection methodologies focused on identifying patterns or instances that deviate significantly from the norm within a given dataset. The primary objective is to highlight unusual or unexpected behaviours that may indicate potential issues, irregularities, or outliers. Unlike traditional detection scenarios where models are trained on both normal and abnormal instances, anomaly detection often involves training models on normal instances, treating anomalies as rare occurrences. In the context of

anomaly detection, the model learns the patterns of the normal class during training and then aims to detect instances that do not conform to these learned patterns. This approach is particularly useful in scenarios where anomalies are rare and might represent critical events, such as seizure epochs. The ultimate objective is to implement models that can accurately and efficiently identify specific events, which will improve decision-making and problem-solving across a range of applications, such as signal processing, security, and healthcare, because of their subtle nature.

1.3 Seizure Detection Methods

Although often questioned, the two-states model in epilepsy (i.e. seizures and seizures-free periods) is a concept generally accepted [1]. While ictal period refers to a seizure event, the interictal period refers the state between seizures (seizure-free, or non-seizure period). In the last decades, some groups also consider a preictal period, i.e. the state immediately before the epileptic seizure [1]. Seizures are a major challenge for many people as they occur randomly and can be dangerous, leading to falls, violent movements, confusion and SUDEP. Seizure detection devices have been proposed as a way to lower the risks related to seizures [15] and to monitor them more efficiently.

Both at home and in a hospital, seizure detection devices can be a valuable tool for epilepsy monitoring. Effective detectors can alert caregivers and alleviate the distress of those suffering from nocturnal seizures by reducing the feeling of helplessness at night [18]. Non-invasive, precise, affordable, user-friendly technologies that are suitable for home and clinical monitoring environments are desperately needed.

Current seizure detection technologies include EEG-based detectors (scalp or implanted EEG recordings), autonomic activity change detectors (electrodermal response, heart rate), and movement detectors (accelerometers, bed alarms, muscular activity, video-based motion features) [19]. However, the effectiveness of each of these detectors varies and is reliant on the particular type of seizure experienced by the patient. For example, alterations in movement are readily apparent only during seizures that involve a significant motor component [19].

Since the gold standard for seizure detection involves using EEG, several methods have been published in the literature that perform automatic detection of seizures by analysing EEG signals. These methods have shown to achieve good results, with sensitivities in the 90% range [20]. Alternative sensing modalities have been proposed to assist with the task of long-term seizure monitoring and detection. These focus on the seizure manifestations rather than the physiological signals generated at the origin of the seizure [20].

In exploring alternative methodologies for seizure detection beyond the traditional EEG-based approaches, a diverse array of methods has been implemented, demonstrating the growing number of technologies aiming to improve real-time monitoring, accuracy, and accessibility in the management of epilepsy.

Video electroencephalography (VEEG)

The typical procedure for detecting and identifying epileptic seizures involves examining brain activity. This is commonly done through using electroencephalography (EEG), which entails sensing the electrical signals generated by the brain using multiple electrodes placed at specific points on the scalp. Afterward, the data is analysed and annotated to identify biomarkers of epileptic activity [20].

Video-EEG monitoring, in which both the video and EEG signals are recorded, is the gold standard for seizure detection in clinical settings. Typically, this is carried out at hospitals or specialized clinics under the direction of qualified clinical personnel. Indeed Video-EEG can provide information on both the brain activity and the physical symptoms of seizures, but it necessitates hospitalization for monitoring. EEG-based systems for epilepsy monitoring need a significant number of electrodes, from 8 to 64, which are typically applied to the patient's scalp during to record the brain activities. Since the location of the seizure's source is typically unknown, the large number of electrodes is crucial [21]. Further, the method's complexity and cost make it impractical for widespread implementation as a tool that most patients may use at home for long-term monitoring to support long-term disease management [20].

In addition to the in-hospital monitoring, patients frequently have to keep diaries at home to record their seizures occurrences. Although useful, this can be highly untrustworthy. Patients' awareness or memories of seizure events depend on a variety of subjective criteria, including their mood, perception, duration, trigger, and intensity [22]. Moreover, one of the main issues that patients must deal with is the underreporting of seizures, which can lead to a variety of issues with diagnosis, treatment, and clinical management [16]. Consequently, there is a push to improve the at-home care for epilepsy patients to help with the diagnosis and management of their condition through the use of long-term monitoring systems.

Accelerometer-based seizure detection systems

Accelerometer-based seizure detection systems have been developed; these systems typically use sensors placed on the arms and/or legs to identify the characteristic motions of seizures. There is a large range of detection sensitivities, from 16% to 100%, in studies that use accelerometery as the sensing modality. The fact that accelerometer-based systems depend upon the movements of the body area where the sensor is placed during seizures contributes to this wide range of sensitivities. This may differ based on the type of seizure. Thus, generalized tonic clonic seizures (GTCS) are best treated with these systems. The low false positive rate (FPR) suggests that these technologies are highly selective and effective in identifying seizures involving motor components [20].

Electromyography (EMG)- based seizure detection systems

Systems based on electromyography (EMG), which detect changes in a particular muscle's activity brought on by seizures, have also been created. Depending on which body parts are most active during seizures, the sensors are often positioned in areas like the arms or legs. The disadvantages of accelerometer-based and EMG-based devices are comparable. Due to the movement-based nature of this seizure detection modality, the number of sensors, the type of seizures, and the position of the sensors with respect to the moving part of the body all affect how well systems based on it operate. Sensitivity values of EMG detection systems are between 57% and 95%. The most sensitive EMG systems, however, also depend on an extremely laborious sensing setup that includes the installation and continuous control of the electrode’s impedance and contact with the skin. The FPR reported in EMG-based research is low [20].

Changes in heart rate variability

Patients may have variations in cardiac activity, which show up as changes in heart rate variability, during or after seizure events [11]. This has also been investigated as a possible physiological method for seizure detection. Sensitivities ranging from 57% to 100% were achieved in studies of seizure detection algorithms using heart rates from ECG recordings. However, those studies were limited to patients who were known to have marked ictal autonomic changes beforehand and depended on patient-specific cut-offs [20].

Photoplethysmography (PPG)-based seizure detection systems

PPG has also been used to detect blood volume changes in the microvascular bed of tissue. But in this case, the sensitivity was either significantly reduced when compared to using ECG, or it relied on signals in which sections with movement artefacts had been manually removed [20].

Bed movement-based seizure detection systems

Non-contact bed movement sensing is a sensing technique that has been employed to automatically detect and record seizures while a person is asleep. Sensitivities ranging from 2.2% to 89% have been documented. The various seizure types that occur in the evaluation datasets can account for these inconsistencies [20].

Video-based seizure detection systems

Video recordings has also been investigated for non-contact seizure detection. Although these methods appear to work well for convulsive seizures, their efficacy significantly declines for other motor seizures or seizures with mild motor symptoms [20].

Electrodermal activity (EDA)-based seizure detection systems

Another sensory modality that has been utilized for seizure detection is the electrodermal activity. During a seizure, the nervous autonomic reaction may result in increased sweating, which lowers skin resistance. This method was tested on tonic-clonic seizures, and it demonstrated good sensitivity (92% to 95%) [20].

Respiration changes-based seizure detection systems

Since seizures often cause changes in breathing, monitoring is crucial in preventing SUDEP. It can also be used to identify sighs, yawns, and other signs of arousal. Arousal and gasping are two key processes in auto-resuscitation, and low arousability may indicate near-SUDEP. Many techniques are available for breathing monitoring, such as vibration signals from the larynx's turbulence during breathing or transcutaneous audio. Other research concentrated on a miniature device that can monitor airflow by listening for sound produced by turbulence in the human respiratory system and attaching it to the skin of the suprasternal notch in the neck [23].

1.4 Rationale for Audio-Based Seizure Detection

Audio-based seizure detection offers several advantages over traditional methods, making it a promising approach in the field of epilepsy management. Traditional methods often rely on intrusive or impractical monitoring devices, limiting continuous and unobtrusive seizure detection. Medical devices that patients wear, such as an accelerometer wristband or an EEG, are considered intrusive tools. Non-intrusive methods include video surveillance and audio recording [24]. Thanks to recent developments in machine learning and audio technologies, a helpful substitute for non-contact epileptic patient monitoring is based on the use of audio recordings [25]. A key advantage is that they can passively monitor seizure activity without causing discomfort or inconvenience to the individual. Invasive devices, on one hand, can carefully track important data, but on the other hand, some hurt the patients and others can come loose accidentally or as the result of muscular spasms [24]. For example, while EEG-based seizure detectors are effective for a wide variety of seizure types, long-term home monitoring is not feasible with them. Audio and video recordings are always accessible. However, they are not as precise as the intrusive methods [24].

One particular benefit of non-invasive devices is that they are cost-effective [24] compared to other sensor-based systems, as they often require less specialized equipment and infrastructure. In many cases, audio sensors are relatively affordable and can be easily deployed in various settings, such as at home, reducing the overall costs by eliminating the need for hospital admissions and clinical observations [23]. In contrast, invasive monitoring needs hospital stays, which can be costly over time and result in extensive waiting lists for patients [24].

Audio recordings offer a widely accessible means to capture physiological changes during seizures. The distinctive acoustic signatures associated with seizures present an opportunity for leveraging sound-based information in the development of efficient and user-friendly seizure detection systems.

Epileptic patients produce a variety of sounds during seizure events that could potentially be detected using audio-based methods. These sounds can vary depending on the specific characteristics of the seizure and the individual. Some typical sounds associated with these types of epilepsy include:

1. Vocalizations: This may include screams, cries, or other vocal expressions resulting from involuntary movements or altered consciousness during seizures. They are typically observed during tonic-clonic seizures, whereas stereotyped vocalizations are associated with some focal seizures.

2. Motor movements: Seizures can involve motor symptoms such as jerking movements, thrashing, or repetitive actions, which may generate corresponding sounds such as banging or thumping. Also, bed-related sounds are caused by the motor component of the seizure.
3. Breathing patterns: Changes in breathing patterns, such as gasping, irregular breathing, or periods of apnea (temporary cessation of breathing), may occur during seizures and produce distinct sounds.
4. Vocal automatisms: Some individuals may exhibit automatic vocalizations or repetitive phrases known as vocal automatisms during seizures, which can manifest as nonsensical speech or repetitive sounds. Laughs for instance, are characteristics of "gelastic" seizures.
5. Environmental interactions: Seizure activity may lead to inadvertent interactions with the environment, such as knocking objects over, bumping into furniture, or rustling noises caused by movements.
6. Altered speech: In cases where seizures affect language centres in the brain, individuals may experience disruptions in speech production or articulation, resulting in slurred speech or distorted sounds.

Detecting and analysing these types of sounds using audio-based technologies, such as microphones or wearable devices, could offer valuable insights into seizure activity and aid in the development of automated seizure detection systems. By capturing and analysing sound patterns, it may be possible to improve the accuracy and efficiency of epilepsy monitoring and provide timely interventions for individuals experiencing seizures. Detecting seizures can alert caregivers to take action, especially during night-time monitoring [23].

1.5 State of the art: audio-based seizure detection

Few studies have been conducted on audio-based seizure detection. Moreover, differently from this thesis, the majority of studies found in literature focused on generalized tonic-clonic seizures and not on focal seizures. However, non-contact audio-based seizure detection methods have been relatively explored. These systems work by listening for noises in the surrounding environment that are linked to seizures, as those made when someone rolls over on top of a mattress or screams or produce any kind of noise both vocally and physically. The reported sensitivities, which ranged from 4.3% to 62.5%, are significantly lower than those obtained from other sensing modalities [20].

Van Andel et al. implemented multimodal detection methods using accelerometry (ACM), electrocardiogram (ECG), video and audio signals [26]. They monitored 43 patients and 23 of them had epileptic seizures during the recordings, for a total number of 86 generalized tonic-clonic seizures. The devices employed for this research were not fully contactless, ACM detectors were attached to the upper arm and 2 wired electrodes on the chest were connected to the arm to record ECG signals. This study revealed a sensitivity ranging from 72% to 87%.
Carlson et al. conducted an audio-based seizure detection study in the United States with 64 pediatric and adult patients, for a total number of 8 generalized tonic-clonic seizures. The recording device was positioned between bed and mattress and obtained a sensitivity of 63% [27].
Another study under the guidance of Fulton et al. utilized a mixture of audio signals and surface pressure. In this research, 27 pediatric patients and 15 instances of generalized tonic-clonic seizures were examined. The utilization of a device positioned between the bed and mattress revealed a detection rate of 4.3% [28].

A few studies have been conducted using only audio signals to detect generalized tonic-clonic seizures. For example, in one study Arends et al. focused on 10 patients with symptomatic generalized or multifocal seizures and obtained a sensitivity of 81% [29]. In another study conducted by Shum et al., 166 clips of 30 seconds of duration from 83 patients were generated. For each patient, they collected one clip of seizure and one clip of non-seizure as control. This study only proved that epileptologists can accurately identify certain seizure types from audio recordings [25].

1.5.1 Feature extraction

One major issue in building an audio recognition system is the choice of proper signal features that are likely to result in effective discrimination [30]. Features extracted in this study will be further discussed in Chapter 2.

Time-domain features play a crucial role in capturing the temporal characteristics of audio signals during seizures. These features are extracted to discern patterns in the amplitude fluctuations of sounds recorded during seizures. They provide valuable insights into the dynamic nature of audio signals during epileptic events. The time-domain features found in literature and employed for sound-related studies are zero crossing rate and root mean square energy [31].

Time-frequency domain features play a crucial role in audio-based detection methods, particularly in tasks such as seizure detection. By analysing audio recordings in both the time and frequency domains simultaneously, these features provide valuable insights into the temporal and spectral modifications of the audio signal. Time-frequency domain features capture variations in signal energy, frequency content, and temporal dynamics over time, allowing for the identification of key patterns and anomalies associated with seizure events. Common time-frequency domain features used for seizure detection include spectrograms, Mel-frequency cepstral coefficients (MFCCs), and short-time Fourier transforms (STFTs) [32]. Other spectral features based on the STFT currently employed for audio-based detection are spectral roll-off point, spectral centroid and spectral entropy [33], [34].

Analysing the spectral content of audio signals enhances the discriminative power of seizure detection systems. These features enable the extraction of relevant information from audio signals, facilitating the development of detection algorithms for applications in healthcare and beyond.

1.5.2 Machine learning

At the forefront of technological advancement, machine learning represents an important shift in the way computers can learn and make decisions. Fundamentally, machine learning is a branch of artificial intelligence (AI) that focuses on creating models and algorithms that can recognize patterns in data and use that information to predict or execute decisions. This discipline includes a wide variety of methods and strategies, each intended to address particular problems in different fields.

In traditional programming, tasks are carried out by explicit instructions given by human developers. Machine learning, on the other hand, makes use of data-driven approaches to allow systems to recognize patterns and rules on their own. For a model to generalize and make predictions or classifications on new, unknown data, it must first be trained on previous data. This process is known as learning.

Supervised learning and unsupervised learning are the two main categories which machine learning can be divided into. The aim of supervised learning is to classify or label new data based on the patterns and relationships learned from previous data, where the labels were known or provided. Supervised learning, which includes well-known techniques like nearest-neighbour classifiers and linear regression, is the most popular category in machine learning [35]. By identifying patterns and regulating parameters in the training data for which labels (or

desired output) are known, the model gains the ability to predict outputs for new and unseen inputs.

In unsupervised learning, the model is not provided with labels. Instead, its goal is to find meaningful or useful patterns within the data. Data visualization and exploratory data analysis are only two of the numerous applications for this method. Clustering analysis, such as K-means, is an illustration of unsupervised learning [35].

Machine learning has found numerous applications in the field of biomedical engineering, transforming the way that complex biological data is analysed and understood. One prominent area of application is in medical image analysis, where machine learning techniques are used for tasks such as image segmentation, feature extraction, and disease diagnosis. For example, convolutional neural networks (CNNs) have been highly successful in automatically detecting and classifying abnormalities in medical images such as MRI scans, X-rays, and histopathology slides [36].

Machine learning also plays a crucial role in personalized medicine by analysing patient data to predict disease outcomes, treatment responses, and optimal therapeutic strategies. This includes the development of predictive models for patient risk stratification, drug discovery, and designing personalized treatment plans based on individual patient characteristics.

Machine learning techniques are currently applied in wearable and implantable medical devices for real-time monitoring of physiological signals, early detection of health abnormalities, and facilitating remote patient monitoring. These devices use predictive analytics, signal processing, and anomaly detection algorithms to give patients quick interventions and useful information.

Overall, the integration of machine learning with biomedical engineering has the potential to significantly advance our understanding of human health and disease, leading to more effective diagnostic tools, treatments, and healthcare interventions.

Despite its transformative potential, machine learning poses challenges, including ethical considerations, interpretability of complex models, and concerns related to bias in algorithmic decision-making. In fact, the complex nature of machine learning models and their training algorithms frequently make it challenging to ensure their efficacy and may limit the understanding of the models. Moreover, ML models may need a substantial amount of training data [35]. The success of machine learning is intricately linked to the availability of datasets and advancements in computational power [28].

Achieving a high degree of generalization in sound detection tasks is difficult for a number of reasons. Examples within the same class may differ greatly from one another (intra-class

variability) due to high levels of ambient noise and the multi-source nature of the recordings [35]. Machine learning in acoustics has made significant progress in recent years [35]. Most computational analysis systems dealing with realistic sounds are based on supervised machine learning techniques, which train the system using labelled examples of sounds from each target sound type. In order for the models to be trained, the system developer must set up a collection of plausible scenes (such a street, house, or office) or events (like a passing car, footfall, or dog barking) categories, classes, and provide a sufficient number of labelled samples. Moreover, the supervised learning approach requires reference annotations, which may take different forms, depending on the method of acquisition. Temporal information, such as the start and end timings of sounds that correspond to each class considered for the particular study, should be included in the annotations [32].

1.6 Objective of the study

The aim of the research is to investigate the feasibility of detecting epileptic seizures from audio recordings and to lay the groundwork for future advancements in this area. By utilizing signal pre-processing techniques to extract relevant features from audio data, such as spectral characteristics, temporal patterns, and frequency distributions, the study seeks to train machine learning algorithms to accurately classify audio segments corresponding to epileptic seizures and non-seizure periods. Additionally, the study aims to explore the potential for audio-based seizure detection as a non-invasive and cost-effective monitoring approach by identifying potential limitations and challenges associated with it, as well as suggesting possible future works and developments in the field.

Ultimately, the goal is to advance the field of epilepsy monitoring and management by leveraging audio recordings as a valuable source of information for seizure detection, thereby improving patient care. This includes assisting healthcare professionals, in hospital settings, and caregivers, at home, in the timely identification and management of epilepsy, therefore improving patients' quality of life.

The primary objective of this thesis is therefore to investigate the feasibility and efficacy of using audio recordings for the detection of epileptic seizures. By exploring the unique characteristics of sound during seizures, we aim to contribute to the development of non-invasive and accessible technologies that enhance the accuracy and timeliness of seizure detection.

Figure 3 shows a visual representation of the proposed approach for the detection of epileptic seizures.

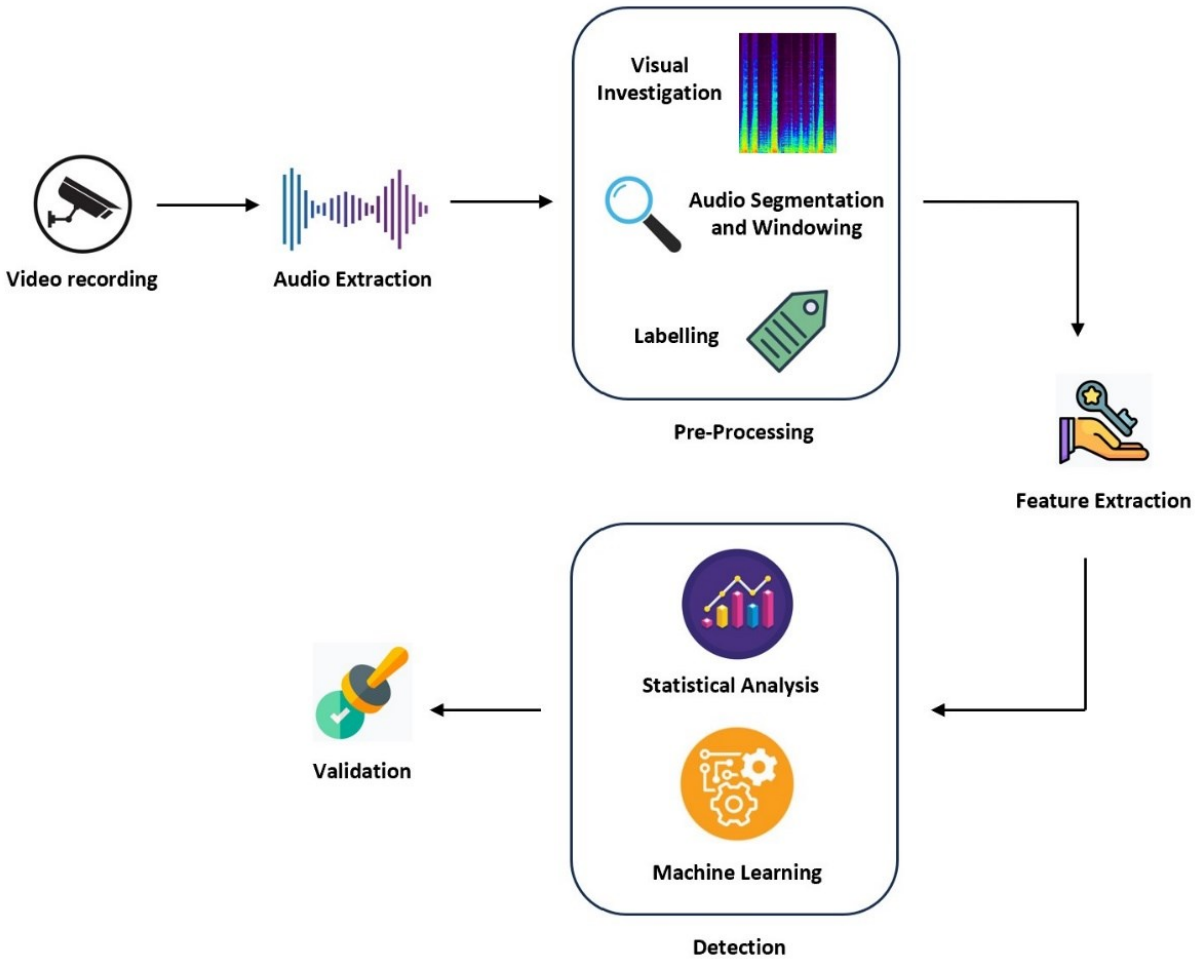

Figure 3. Pipeline of the proposed approach for the detection of epileptic seizures from audio recordings. From the video recordings, we extracted the corresponding audio recordings, then we performed pre-processing steps such as visual investigation, audio segmentation and windowing and labelling of windows. We extracted features from each segment, performed the detection task employing machine learning models and finally, we validated the models.

The thesis is organized as follows:

- Chapter 2 - "Materials and Methods" - contains details about the methodology employed in this study, including the collection and preprocessing of audio data, feature extraction techniques, and the design of the seizure detection models.

- Chapter 3 - "Experimental Results" - presents the results of our analysis, evaluating the performance of the proposed seizure detection system on the dataset.

- Chapter 4 - "Discussion" - presents a discussion about the implications of the results, and it addresses potential limitations.

- Chapter 5 – “Conclusions” - the thesis is concluded by summarizing key findings and outlining directions for future research in the field of audio-based seizure detection.

By carrying out this investigation, we hope to provide knowledge to improve seizure detection technology, with a focus on the special benefits provided by audio recordings.

CHAPTER 2: MATERIALS AND METHODS

2.1 A word about the hosting laboratory

The Paris Brain Institute, also known as ICM (**Figure 4**) is an international research centre focusing on neuroscience studies. The ICM is a private foundation of public utility located within the La Pitié-Salpêtrière hospital premises, in Paris. La Pitié-Salpêtrière is the largest hospital and university medical centre (CHU) in Europe, and is part of the AP-HP network (Assistance Publique – Hôpitaux de Paris), which comprises 39 hospitals.

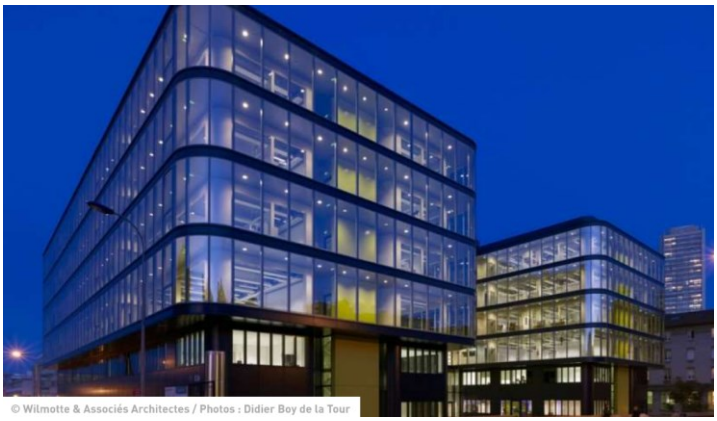

Figure 4. Institut du Cerveau (ICM – Paris Brain Institute).

This research centre is unique because it combines healthcare professionals, patients, and researchers to create therapies more quickly and effectively than they could elsewhere. Because of its creative design and structure, it is a global research centre without comparison. At the institute, more than 700 scientists from many countries and backgrounds work together to conduct cutting-edge research in the neurosciences field. By supplying researchers with equipment throughout 22000 m^2 of laboratory space, 1200 m^2 dedicated to clinical research, and 1000 m^2 for startups, the ICM aims at shortening the gap between research and clinical therapy.

The team "Dynamics of epileptic networks and neuronal excitability", led by M Chavez (PhD, DR-CNRS), Prof. S Charpier (PhD), and Prof V Navarro (PhD, MD), studies the development of epilepsy in the brain (epileptogenesis) and the occurrence of seizures (ictogenesis). The team has a translational and multidisciplinary approach. The team uses in vivo recordings in both epileptic patients and animal models where seizures are induced to

better understand epilepsy and its underlying mechanisms. The aim is to detect and predict epileptic seizures through various approaches.

2.2 Data Collection

At the renowned medical institution, “Hôpital Universitaire Pitié-Salpêtrière” located in Paris, a dedicated initiative is in progress, aiming to dive into the complexities of epilepsy management. Specifically, pharmacoresistant patients, candidate to a surgery, undergo meticulous intracranial electrode implantation procedures. These procedures are planned to comprehensively study brain activity during seizure episodes to have a better understanding of where the neurons involved in the seizure are located.

The observational phase spans over two to three weeks and is designed to capture a comprehensive understanding of the patients’ neurological dynamics. Throughout this period, a continuous and vigilant monitoring approach is implemented, where the patients and their immediate surroundings are documented through Video Electroencephalography (VEEG) recording technology. This approach not only provides an opportunity to analyse seizures in details but also provides us with valuable resources for this research project centred on the detection of seizures from audio recordings.

2.2.1 Instrumentation

Within the context of this thesis, only videos and their audios, obtained by VEEG technology were considered. The instrumentation employed to record the videos, as shown in **Figure 5**, include the following camera and microphone: AXIS M5525–E PTZ Network Camera and AXIS T8351 Mk II Microphone 3.5 mm. The AXIS T8351 Mk II 3.5mm device is a low-noise analog hemispherical microphone, with a high sound [37].

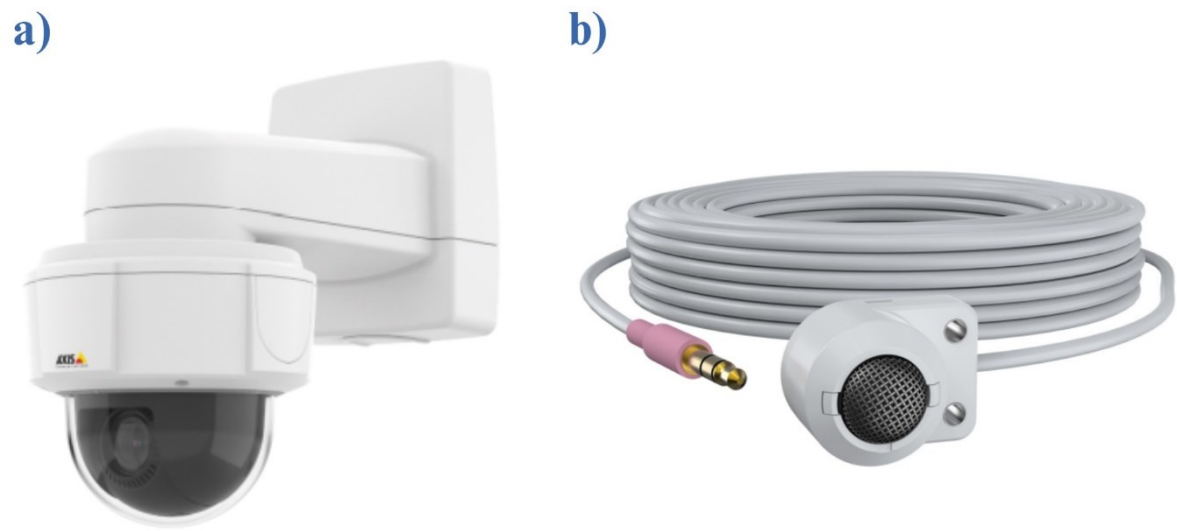

Figure 5. a) AXIS M5525–E PTZ Network Camera with support to install it on the wall of the patient's room and b) AXIS T8351 Mk II Microphone 3.5 mm.

2.2.2 Video Selection

The long-term video-EEG recordings were reviewed by a clinical neurologist. In particular, through visual analysis and EEG analysis of hours documentation, the neurologist selected 5 patients and 15 videos for a total of roughly 400 minutes (nearly 7 hours) of video recordings. The neurologist then annotated the specific time in which the seizures occurred and provided additional details about the type of seizure (focal and focal to bilateral), the patient's state of vigilance and if the seizure occurred either during the day or during the night, as indicated in **Table 1**.

Patient	Seizure Start	Day Period	State of Vigilance	Seizure Type	Audio Duration
3105	06:00:34	nighttime	sleep	focal	06:47
3132	07:16:42	nighttime	sleep	focal	02:56
3152	03:39:50	nighttime	sleep	focal to bilateral	02:23:41
3249	16:46:57	daytime	awake	focal	01:59:57
3300	20:23:43	daytime	awake	focal to bilateral	01:59:57

Table 1. ID of patients, time of the day in which the seizures start, day period in which the seizures occur, state of vigilance of the patients, and seizure types, and overall duration of the audio recordings.

Each one of these videos showed patients in their hospital room and a variety of sounds were produced. Some videos contained periods of silence, periods of TV sounds or speech, as well as external voices when other people, such as nurses, doctors or caregivers were talking close by the room. Finally, we were provided with one clear ictal period for each subject to use for

this thesis work. The ictal period, or seizure epoch, is limited between two markers and it can vary across patients.

A representation of patients undergoing video-electroencephalogram (EEG) set-up is shown in **Figure 6** and **Figure 7**, which capture their experiences during nighttime and daytime sessions. By using video recordings of patients, clinicians have a contextual perspective of their immediate surroundings. These figures provide a look into the monitoring environment of the patients' room at the epilepsy unit.

Figure 6. A patient in their hospital room at Pitié-Salpêtrière, during the daytime, undergoing a VEEG procedure.

Figure 7. A patient in their hospital room at Pitié-Salpêtrière, during the nighttime, undergoing a VEEG procedure.

2.3 Audio extraction from video recordings

Because this study aims to detect epileptic seizures from audio recordings, the first step of the analysis was to extract audio recordings from the patients' video footage. Different functions from the Audio Toolbox™ of MATLAB R2023b were employed to import audio files. Original audio-video were in MP4 format. Sampling rate of audio recordings was of 32 kHz with a bit rate of 62.24 kbps.

2.4 Exploratory Data Analysis

In this section we embark on an examination of the characteristics of audio signals, utilizing visualization tools such as the waveform, spectrogram, and Mel spectrogram (discussed in detail later). This introductory phase plays a crucial role in identifying patterns, variations, and peculiarities in the data, providing a visual overview of the underlying dynamics.

Through the visualization of the waveform, we capture the temporal variations of the signals, while the spectrogram offers a graphical representation of the spectral distribution in the frequency domain. Simultaneously, the Mel spectrogram provides an even more refined perspective, emphasizing features perceived by the human ear. These visual tools not only facilitate the understanding of the peculiarities of audio signals but also serve as a guide in the subsequent phase of segmentation, windowing, labelling, and feature extraction (explained below). The result of these pre-processing steps will constitute our dataset, ready to undergo machine learning procedures.

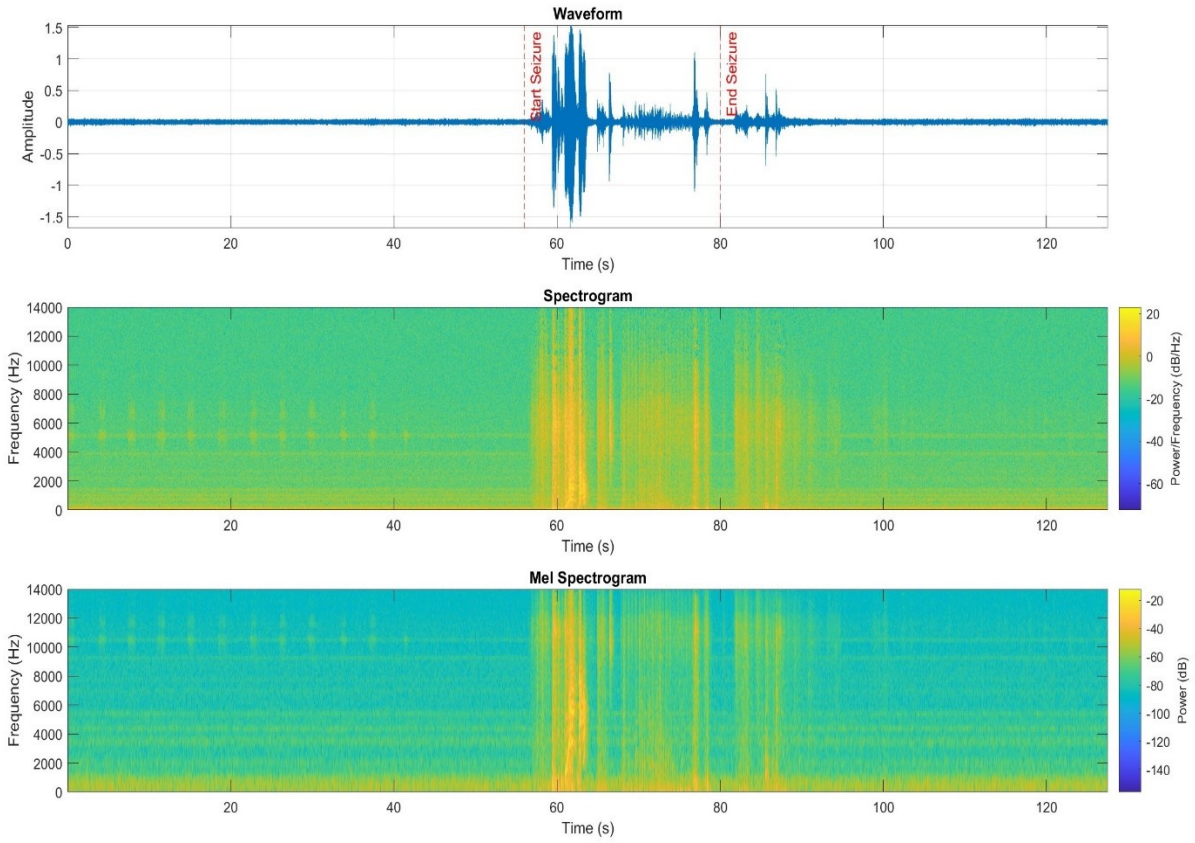

Figure 8. Plot of waveform, spectrogram and Mel spectrogram of video 261 of patient 3132 with seizure epoch markers. Seizure occurs between the second 56 and the second 80.

The waveform of an audio signal illustrates how the signal's amplitude changes over time. It is a basic graphical representation that shows how the sound wave's pressure levels change as it develops. When it comes to waveform analysis, the vertical axis represents the sound wave's amplitude or intensity, while the horizontal axis indicates time. Whereas the natural units of the analog microphone's output are Volts, the different decoding techniques used to extract the

audio signal from the MP4 file yield a normalized signal, with values between -1 and 1. Nevertheless, as indicated in the technical page of MATLAB, this signal might exceed such values. Units of the extracted audio from MP4 files are therefore dimensionless.

Through close examination of the waveform, one can identify the subtle aspects of the sound, such as periods of silence, the onset and offset of distinct sounds, and the overall dynamics of the auditory experience. In the top plot of **Figure 8**, for instance, one can clearly observe a sound event corresponding to an ictal period of the patient. In audio signal processing, this visual representation is a fundamental tool that facilitates a qualitative comprehension of the patterns and behaviours incorporated into the sound data.

The spectrogram of an audio signal is a powerful representation that provides a detailed insight into the frequency content of the signal over time. A spectrogram is obtained by applying the Short-Time Fourier Transform (STFT) to the signal, previously broken into short overlapping segment, and computing the Fourier transform. Mathematically, the STFT of a signal $x(t)$ at a time t and frequency ω is given by:

$$X(t,\omega) = \int_{-\infty}^{+\infty} x(\tau) \cdot \omega\,(t-\tau) \cdot e^{-j\omega\,\tau}\, d\tau \qquad (1)$$

where:

- $X(t,\omega)$ is the STFT of the signal at time t and frequency ω
- $x(t)$ is the input signal
- $\omega(t)$ is a window function that defines the length and shape of the segment
- τ is the time variable for integration
- $e^{-j\omega\,\tau}$ represents the complex exponential term that provides the frequency information

As seen in the middle plot of **Figure 8**, the spectrogram shows the evolution of the intensity, or "loudness," at different frequencies during the development of the ictal event. The vertical axis displays frequencies, while the horizontal axis corresponds to time units [33].

When a spectrogram is generated, several parameters such as window type $\omega(t)$, length, and overlap can be modified to get the desired time and frequency resolution and bandwidth of the spectrogram. It is important to keep in mind that while a larger window may provide better frequency resolution, time resolution will drop as a result. On the other hand, although a higher overlap results in a better time resolution, it is computationally more intensive. A frequently used value for the overlap is 50%, which corresponds to half of the window.

The seizure event's unique frequency patterns are displayed in the spectrogram, making it possible to distinguish between the epileptic and non-seizure periods. The time-frequency representation in the spectrogram effectively localizes the onset, duration, and conclusion of the seizure within the audio signal, providing valuable temporal information for analysis. Observable changes in the spectral content during the seizure reveal specific frequency shifts, modulations, or intensity variations that are characteristic of epileptic activity, further enhancing the localization analysis of the audio signal. The visually discernible patterns may serve as a foundation for the development of automated seizure detection algorithms, leveraging the distinctive features captured by the spectrogram for accurate and efficient identification of seizure events in audio signals.

The Mel spectrogram is a time-frequency representation of an audio signal that is obtained by applying a Mel filter-bank to the short-time Fourier transform (STFT) magnitude spectrum of the signal. The Mel scale is a perceptual scale of pitch or frequency, and it is based on human auditory perception (non-linear), where equal distances on the Mel scale are perceived as equal pitch intervals.

The estimation of the Mel spectrogram involves the following steps:

1. Compute the STFT of the audio signal

2. The crucial step in obtaining a Mel spectrogram involves transforming the linear frequency scale of the spectrum into the Mel scale, which better reflects the way humans perceive pitch differences. This transformation is achieved by applying a Mel filterbank to the magnitude spectrum obtained from the STFT. The Mel filterbank consists of a set of triangular filters spaced evenly on the Mel scale. Each filter overlaps with adjacent filters and is designed to capture different frequency bands.

3. Multiply the magnitude spectrum obtained from the STFT with each filter in the Mel filterbank.

4. Sum the results of the multiplication for each filter to obtain the Mel spectrogram.

Mathematically, the formula for computing the Mel spectrogram $S_m(t,f)$ at time t and frequency f can be expressed as:

$$S_m(t,f) = \Sigma_{m=1}^{M} |X(t,f) \cdot H_m(f)|^2 \qquad (2)$$

where:

- $X(t,f)$ is the magnitude spectrum of the STFT at time t and frequency f,

- $H_m(f)$ is the frequency response of the m-th Mel filter in the Mel filterbank,

- M is the number of Mel filters in the filterbank.

The Mel spectrogram provides a more perceptually relevant representation of the audio signal, making it particularly useful in tasks such as speech processing, music analysis, and sound recognition. Showing both the standard spectrogram and the Mel spectrogram of an audio signal can provide a more informative and perceptually relevant representation of the signal.

Parameters used to obtain both the spectrogram and the Mel spectrogram are collected in **Table 2**.

	Spectrogram	Mel Spectrogram
Window (samples)	512	512
Overlap (samples)	Window/2	Window/2
FFTLength (samples)	2048	2048
numBands		128

Table 2. Parameters to calculate both the spectrogram and Mel spectrogram. Window length refers here to the segment's length involved in the STFT algorithm.

2.5 Data Pre-Processing

To improve performances of different machine learning models, a few pre-processing steps need to be completed, such as conversion to the frequency domain, audio segmentation and windowing and labelling of windows [33].

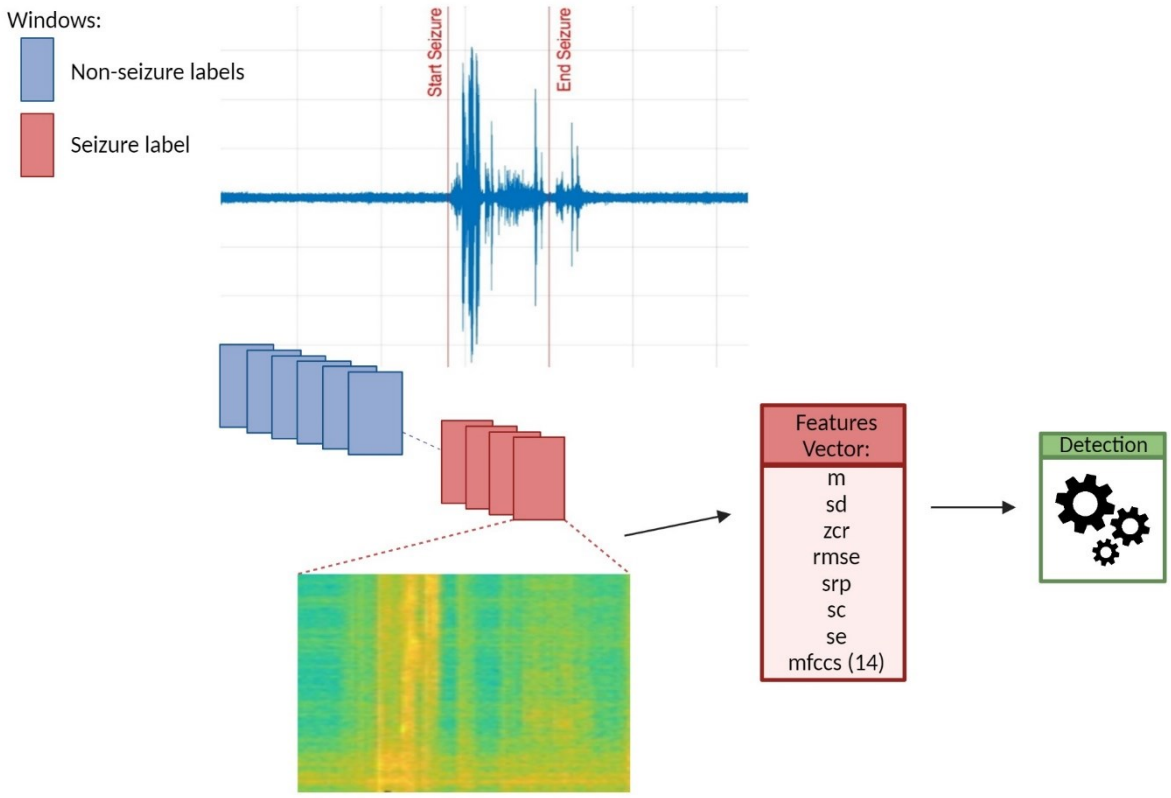

Figure 9. Visual representation of pre-processing steps, features extraction step, creation of the dataset step and detection step.

2.5.1 Audio segmentation and windowing

Audio segmentation and windowing are essential processes in the analysis of audio signals, particularly in the context of signal processing and feature extraction.

Audio segmentation involves dividing a continuous audio signal into shorter, manageable segments, enabling a more focused analysis of distinct events or patterns within the signal. To this study, we divided the audio signals into smaller segments of 10 seconds in length with a 50% overlap.

Windowing, on the other hand, is a technique used to mitigate issues related to the abrupt start and end points of segments by applying a window function to each segment. In our implementation, we utilized the Hamming window function, a commonly used windowing function in signal processing. The Hamming window smoothly tapers the edges of each segment, gradually reducing the amplitude towards the edges while maintaining the centre amplitude. This tapering minimizes spectral leakage and information loss, facilitating more accurate spectral analysis and feature extraction. The choice of an appropriate window function is important, as it influences the balance between temporal and frequency resolutions in

subsequent analyses. With the action of overlapping the audio vector is not simply segmented into consecutive portions. Instead, each new segment includes a portion of the previous segment and a portion of the next one. When short audio events occur at the border of a given segment, overlapping makes sure they are regarded as complete in the following overlapped segment [38]. This can be beneficial to prevent the loss of crucial information at the edges of the segments and to maintain signal continuity [32], [38].

2.5.2 Labelling

After obtaining the segments from each audio recording, the following step involves labelling each window before proceeding with the feature extraction process. The labelling of windows in the context of signal processing, and particularly in audio analysis, involves assigning semantic tags or categories to individual segmented portions of a signal. The labelling process is crucial for annotating each window with meaningful information, such as identifying distinct sound events, delineating between silence and speech, or marking specific characteristics like epileptic seizure. Accurate labelling provides a foundation for supervised learning tasks, facilitating the training of machine learning models for tasks such as sound classification or event detection. The challenge lies in ensuring that the assigned labels effectively capture the inherent features of each window, contributing to the development of robust and context-aware audio analysis systems.

There are several methods for obtaining the labelling of windows. The most common method of getting annotations is manually, that is, designating someone to listen to the audio that is going to be utilized and mark the activities for each class. Appropriately annotating a portion of audio is a laborious task that can easily take much longer than the audio itself. However, for some kinds of sound classes, human annotation is frequently the only way to get labels [32].
To do that, all videos underwent a thorough rewatch. For each patient we created excel tables. Example of patient 3152 is shown in **Table 3**, where we annotated the start time, the end time and the label corresponding to every event occurring during each video. For practical reasons, we also inserted relevant information such as the name of the folder and the name of the videos considered. The employment of an organized approach guarantees the precision and level of detail in the annotated data, while also promoting efficient arrangement and availability for later phases of examination and handling.

start_time	end_time	label	folder	video
0	114	tv	3152	VID_243
115	198	other	3152	VID_243
199	551	tv	3152	VID_243
552	890	other	3152	VID_243
891	1018	tv	3152	VID_243
1019	1034	other	3152	VID_243
1035	1143	silence	3152	VID_243
1144	1198	other	3152	VID_243
1199	1314	movement	3152	VID_243
1315	1358	other	3152	VID_243
1359	3400	resting-silence	3152	VID_243
0	2487	resting-silence	3152	VID_244
2488	2707	seizure	3152	VID_244
2708	3400	resting-silence	3152	VID_244
0	397	resting-silence	3152	VID_245
0	1216	resting-silence	3152	VID_247
0	72	resting-silence	3152	VID_249

Table 3. Table of patient 3152 containing information about the time of interest of each sound event appearing in each video, its corresponding label, the name of the videos and the folder where the videos are.

Some of the labels that can be assigned to the windows of audio signals are: "seizure," "resting-silence," "movement," "TV," "extvoice," (referring here external voices) "speech," "phone," "door," "bird," "shower," or "other". Particularly, the last one was used when sound identification posed a challenge. We decided to label different sound events occurring within every video in a meticulous manner in order to make the dataset ready and available for future studies of sound classification. Nevertheless, given the reduced size of the dataset, in this thesis research we grouped all sounds into two groups: the "seizure" and "non-seizure" categories.

2.6 Features Extraction

The goal of feature extraction is to gather data to identify or categorize the target sounds, which will reduce the computational cost and simplify the subsequent modelling stage [32].

Feature extraction is a critical step in the analysis of audio signals, playing a pivotal role in transforming raw waveform data into a set of meaningful and representative features. To represent audio in a clear and non-redundant manner, audio analysis typically relies on acoustic features that are extracted from audio signals [32]. A low intra-class variability of features (i.e. between features extracted from samples allocated to the same class) and high variability between features extracted from samples assigned to different classes (high inter-class

variability) are prerequisites for recognition algorithms [39]. The purpose of feature extraction is to convert the signal into a representation that optimizes the classification algorithm's performance in sound recognition. Acoustic features that are pertinent to machine learning are generally related to physical characteristics of audio signal, and are numerical attributes that represent its signal energy, frequency distribution of power, and time variation [32].

In the pursuit of detecting epileptic seizures from audio recordings, a diverse range of features has been extracted to encapsulate various aspects of the signal's characteristics [19]. These features encompass statistical measures such as mean, standard deviation of the signal's amplitude, and zero-crossing rate, offering insights into the signal's central tendency and variability. Additionally, spectral features like spectral centroid and Mel-frequency cepstral coefficients (MFCCs) provide information about the distribution of frequencies, capturing nuances relevant to seizure events.

Since the magnitude of the frequency components of the signal may quickly vary with time, audio signals are typically non-stationary. As a result, the feature extraction method makes use of the short-time processing approach, in which the signal is captured in a quasi-stationary condition by analysis conducted in short-time segments. As mentioned above, the audio signals were divided into 10 sec segments with a 50% of overlap [38]. A windowing function was used to smooth the analysis frames in order to prevent sudden changes at the frame boundaries, which could lead to spectral distortions. The features were computed from each 10s segment to construct a feature vector. These vectors served as the input for subsequent statistical analysis and machine learning models. Moreover, compared to directly analysing the audio signal, a compact feature representation uses less memory and processing power [32].

Time-domain features extracted from the 10 seconds segments of the audio signals include [30]:

- Mean and standard deviation of signal's amplitude x(t) provide information about the data's central tendency and variability. Since sound events (seizures) frequently take the form of departures from reference distribution of statistical characteristics of non-seizure epochs, these characteristics are useful in spotting abnormal or rare sound events.

  In the context of audio signal processing, mean and standard deviation are computed to capture localized statistical characteristics. For each segment, denoted by W, containing n data points, the mean $\bar{X}$ is computed using the formula:

$$\overline{X_W} = \frac{\sum_{i=1}^{n} X_{i,W}}{n} \quad (3)$$

Where $X_{i,W}$ represents each individual data point within the specific window. This calculation provides the average amplitude of the signal within that window Simultaneously, the standard deviation σ_w within the same window is calculated using the formula:

$$\sigma_w = \frac{\sqrt{\sum_{i=1}^{n} (X_{i,W} - \overline{X_W})^2}}{n} \quad (4)$$

This equation quantifies the dispersion or variability of the signal within the segment W.

- Zero Crossing Rate (ZCR) is a feature commonly used in audio signal processing to characterize the frequency content and dynamics of a signal [40]. It represents the rate at which the signal changes its sign, indicating the number of times the waveform crosses the zero-amplitude axis within a specified time window.

  The ZCR is computed as follows:

$$ZCR_W = \frac{1}{n-1} \Sigma_{i=1}^{n-1} \left| sign\left(X_{i,W}\right) - sign\left(X_{i+1,W}\right) \right| \quad (5)$$

  where $X_{i,W}$ represents the $i-th$ data point within the specific segment, and $sign(\cdot)$ is the sign function, returning -1 for negative values, 0 for zero, and 1 for positive values. This calculation provides a quantitative measure of how frequently the signal changes direction within the analysed segment, offering valuable insights into its temporal characteristics.

  Extracting the Zero Crossing Rate (ZCR) feature is valuable in the context of seizure detection because seizure events often introduce abrupt alterations in audio characteristics, such as irregularities in speech frequencies, vocalizations, or other audible manifestations. The ZCR feature, by quantifying the rate of zero crossings or changes in signal polarity, becomes a relevant metric to capture these rapid fluctuations in the audio signal associated with seizure activity. Seizures may manifest as sudden and irregular sounds, which can be reflected in an elevated ZCR compared to the ZCR estimated in non-seizure periods.

- Root Mean Square (RMS) energy is a widely used feature in audio signal processing that quantifies the overall energy or amplitude of a signal. It is particularly valuable in capturing the intensity and strength of a sound signal. In the context of seizure detection from audio recordings, RMS becomes relevant in identifying changes in the overall energy level of the audio signal during seizure events.

  The RMS energy (RMS_W) for a specific window is calculated using the formula:

$$RMS_W = \sqrt{\frac{1}{2}\Sigma_{i=1}^{n}\left(X_{i,W}\right)^2} \quad (6)$$

  where $X_{i,W}$ represents each individual data point within the given window. This computation results in a value that represents the magnitude of the audio signal within the localized time frame. The analysis in a sliding window (10s) allows for the tracking of variations in signal energy over time, making RMS a valuable feature for discerning patterns indicative of seizure onset in sound recordings.

The selection of these time-domain features was mainly influenced by computational considerations. They provide meaningful information about the audio signal while being computationally simple to estimate, making them suitable for real-time processing. Together, these features offer a multifaceted representation of the audio signal. Having a combination of statistical, temporal, and energy-related features increases the likelihood of detecting a variety of anomalies.

The combination of statistical and frequency features aims to comprehensively represent the intricate details of the audio signal, facilitating the development of robust models for accurate seizure detection. The spectral features considered in this study are:

- The Spectral Rolloff Point is a feature in audio signal processing that characterizes the frequency content of a signal by identifying the frequency below which a certain percentage of the total spectral energy is contained. It serves as a valuable descriptor for the upper boundary of the signal's frequency distribution. In the context of seizure detection from sound, spectral roll-off can be indicative of changes in the spectral composition during abnormal events.

The Spectral Rolloff Point (SRP_W) is determined by first obtaining the power spectrum of the signal within each segment. The SRP is then calculated as the frequency below which a predetermined percentage (commonly 85% or 95%) of the total spectral energy resides. Mathematically, for a signal *x(t)* with a power spectrum represented by $P(f)$ for frequency f, the SRP is computed as follows:

$$\Sigma_f P(f) < Percentage\ Threshold \times \Sigma_{all\ frequencies} P(f) \qquad (7)$$

This computation results in a frequency value that represents the point below which the specified percentage of the signal's energy is concentrated. Analysing the evolution of the Spectral Rolloff Point provides insights into the frequency components changes of the audio signal, that can discriminate between normal and seizure-related sound patterns.

- Spectral Centroid is a crucial feature in audio signal processing that characterizes the "centre of mass" or average frequency of a signal's power spectrum. It provides valuable insights into the overall tonal characteristics of the audio signal. The Spectral Centroid (SC_W) is calculated by weighing each frequency component in the power spectrum by its amplitude and determining the center of mass of this distribution.

  The Spectral Centroid is computed as follows:

$$SC_W = \frac{\Sigma_f f \cdot P(f)}{\Sigma_f P(f)} \qquad (8)$$

  This formula essentially represents a weighted average of the frequencies in the power spectrum. Higher Spectral Centroid values indicate that the majority of spectral energy is concentrated towards higher frequencies, while lower values suggest a prevalence of energy towards lower frequencies. The analysis of the Spectral Centroid allows for the tracking of changes in the central frequency of the audio signal, making it a valuable feature for discerning alterations in sounds associated with seizure events.

- Spectral Entropy is a feature in audio signal processing that quantifies the spectral complexity or randomness of a signal. It provides information about the distribution of energy across different frequency components. Spectral Entropy (SE_W) is derived from the normalized spectral power distribution within each time window. This feature is

particularly useful in capturing variations in the spectral content of audio signals over time.

The Spectral Entropy is computed as follows:

$$SE_W = -\Sigma_f \frac{P(f)}{\Sigma_f P(f)} \cdot log\left(\frac{P(f)}{\Sigma_f P(f)}\right) \quad (9)$$

This formula calculates the Shannon entropy of the normalized power distribution, measuring the uncertainty or disorder in the spectral content. Higher Spectral Entropy values indicate a more evenly distributed spectral energy, reflecting greater variability, while lower values suggest a more focused or concentrated distribution. Analysing Spectral Entropy provides a dynamic perspective on the changes in spectral complexity within the audio signal, contributing to the identification of sound patterns associated with seizure events.

- Mel-frequency cepstral coefficients (MFCCs) are the most widely used acoustic features in audio signal processing and speech recognition to represent the spectral content of audio signals (MFCCs) [41]. MFCCs capture the distribution of energy across different frequency bands, making them particularly effective for tasks involving sound analysis.

  After computing the Mel spectrogram from the audio signal using Mel filterbanks as previously discussed, the next step involves transforming it into the log Mel spectrogram. This transformation is applied to approximate the non-linear perception of loudness by the human auditory system, resulting in a representation known as the log Mel spectrogram. Then, the log Mel spectrogram is subjected to a Discrete Cosine Transform (DCT) that produces Mel-frequency cepstral coefficients (MFCCs). These coefficients capture essential characteristics of the audio signal's spectral content. The resulting MFCCs represent thus the spectral content of the audio signal in a compact and perceptually relevant manner.

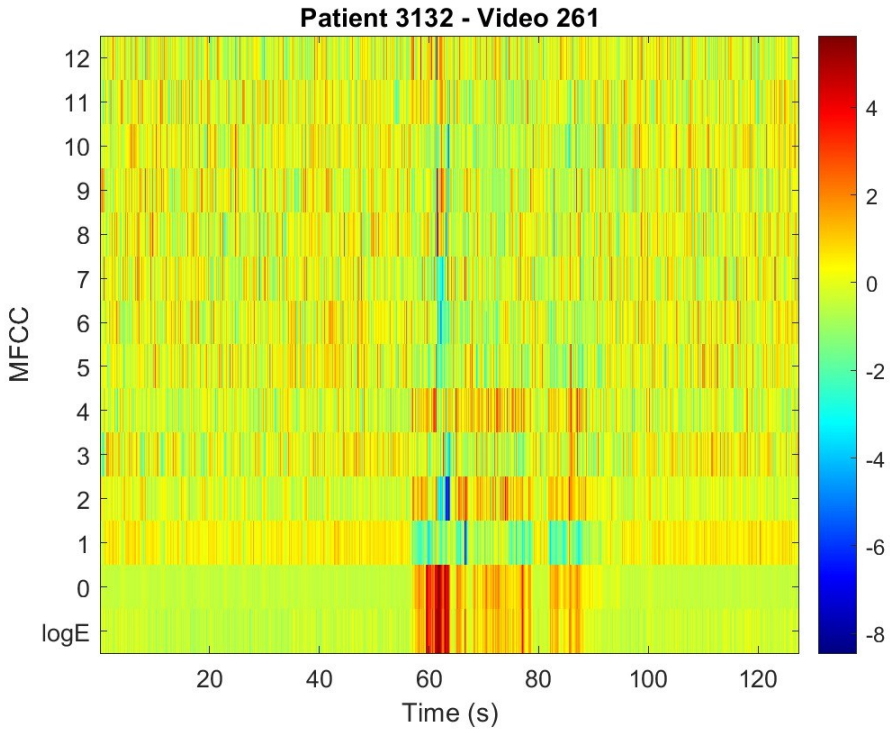

Figure 10. *Visualization of Mel-frequency cepstral coefficients (MFCCs) of video 261 of patient 3132. Ictal period occurs from 56 sec to 80 sec.*

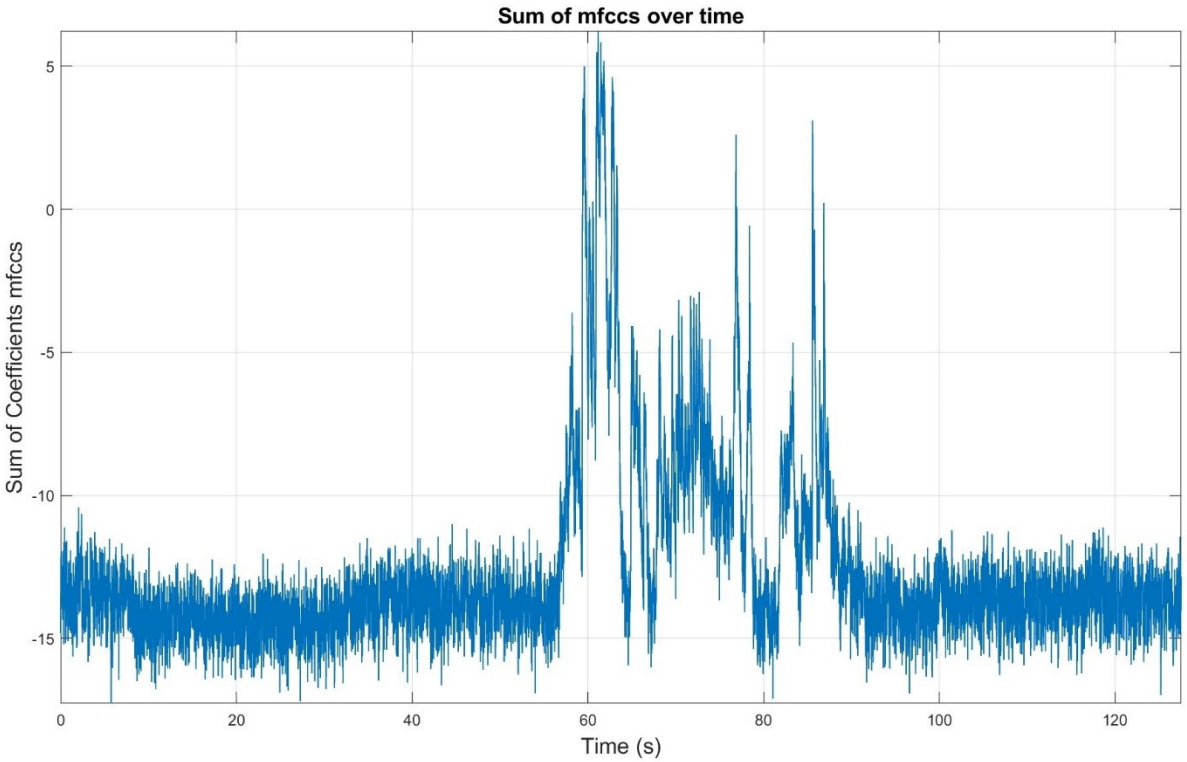

Figure 11. *Visual result of employing the sum of Mel-frequency cepstral coefficients (MFCCs) of video 261 of patient 3132, over time. Ictal period occurs from 56 sec to 80 sec.*

Here, the analysis of MFCCs provides another time-varying representation of the signal's spectral features, making them valuable for discriminating between different sound patterns, including those associated with seizure events.

To illustrate the features described above, **Figure 12** plots the waveform of audio signal obtained from the video 261 of patient 3132, and some features that have been extracted from it, such as: cepstral coefficients, spectral entropy, spectral roll-off point, zero crossing rate and root mean square energy. The red lines represent the two markers for the beginning and the end of the seizure period.

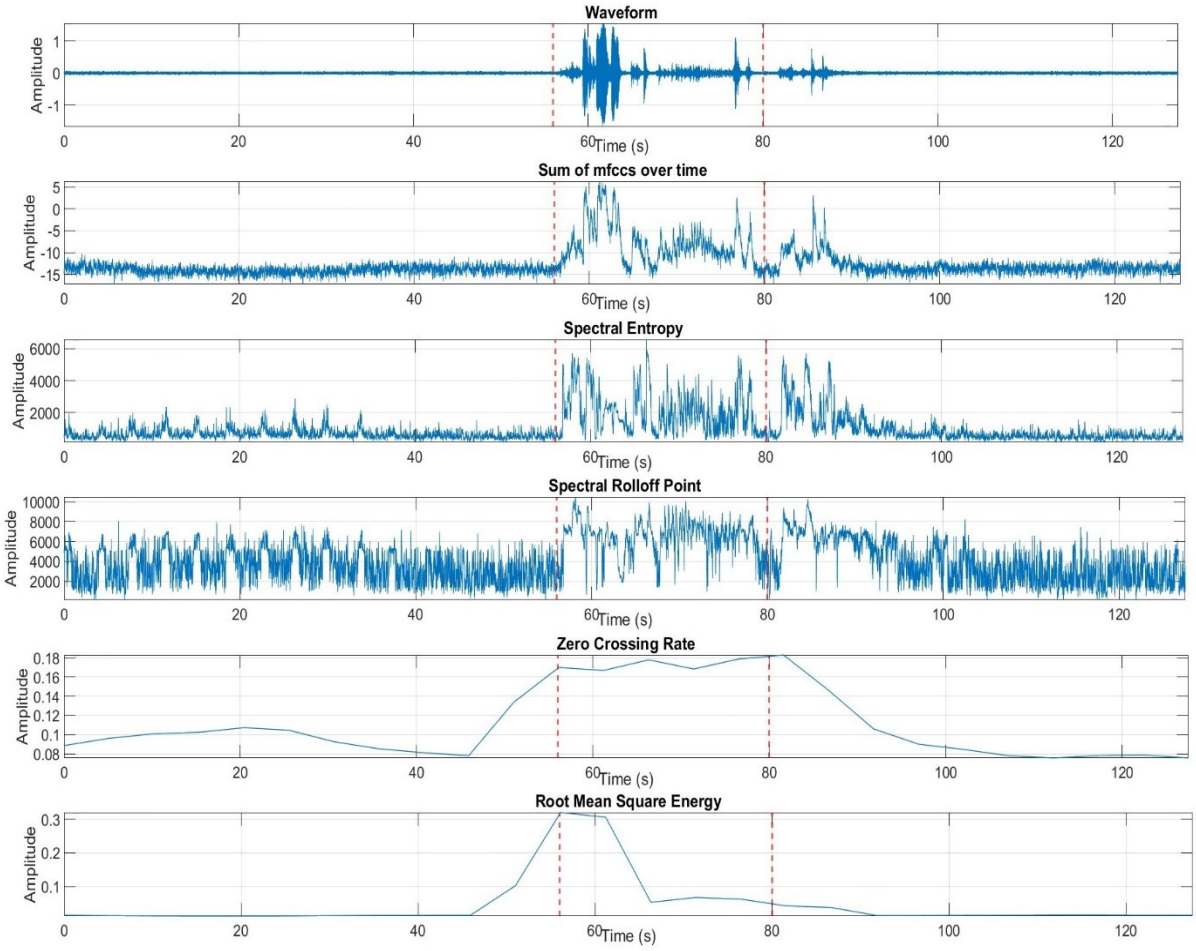

Figure 12. Plot of waveform of audio signal from video 261 of patient 3132 and some of the features extracted from it, such as: mfccs, spectral entropy, spectral rolloff point, zero crossing rate and root mean square energy. In red the markers for the beginning and the end of the seizure period.

2.7 Dataset Characteristics

The dataset employed for this study is structured to capture a comprehensive representation of audio data related to epileptic seizures. Organized on a patient-centric basis, the dataset includes multiple videos for each individual, providing a diverse set of recordings that contains different sounds, ranging from those associated with seizure activity to those representative of non-seizure periods. Within each video, the continuous audio signal is systematically divided into distinct 10 seconds segments (with 50% of overlapping), allowing for a localized analysis of the signal's temporal characteristics (**Table 4**).

Patients	Seizure Periods (s)	#windows seizure	Non-Seizure Periods (s)	#windows non-seizure
3105	120	24	280	56
3132	30	6	150	30
3152	220	44	8400	1680
3249	80	16	7110	1422
3300	650	130	6545	1309

Table 4. *Quantified seizure and non-seizure periods for each patient in terms of both duration (seconds) and the number of windows.*

From each of these windows, a set of features is extracted, incorporating statistical, spectral, and cepstral measures that collectively contribute to a nuanced representation of the audio content.

Specifically, we created two separate sets of features: time-domain features (mean, standard deviation, zero crossing rate and root mean square energy) and frequency domain features (spectral rolloff point, spectral centroid, spectral entropy and mel frequency cepstral coefficients). Time-domain features are a single scalar value per segment of audio, whereas spectral features are a vector per segment of audio. Afterwards, we trained and tested all machine learning models on both sets of features separately.

The reason behind this decision is that dividing features into two groups based on their domain can be useful to improve interpretability and to focus on the most relevant features for each domain, to understand which models are more suited for which feature domain . The decision-making process for this approach was guided by considerations such as feature relevance and interpretability, model complexity and computational efficiency.

The dataset, encapsulated within a MATLAB structure, offers a multifaceted exploration of audio patterns associated with seizures and non-seizure segments, allowing for both patient-specific and collective analysis into the potential auditory markers of epileptic events. As a result, the dataset not only provides a resource for the current study but also creates an adaptable

foundation for future research projects by allowing the addition of more patients and video recordings to the automated process.

2.8 Statistical Analysis

Statistical analysis plays a key role in the detection of epileptic seizures from audio recordings, offering a framework to discern patterns and variations within the data. By employing descriptive statistics from signal, it is possible to gain insights into the distribution of audio characteristics associated with seizures. These investigations provide a basis for the development of detection systems by revealing statistical patterns that differentiate between normal and seizure-related audio. In what follows, all the detection and classification methods were evaluated separately for each video of every patient, whether the epileptic seizure occurred or not.

2.8.1 Mahalanobis distance

Named after the Indian statistician Prasanta Chandra Mahalanobis, this distance metric is particularly valuable when dealing with multivariate data, providing a more robust measure than Euclidean distance in the presence of correlated variables [42]. In the context of audio-based seizure detection, Mahalanobis distance can be employed to assess the similarity of feature vectors extracted from audio recordings to a reference distribution (e.g. the non-seizure periods). By considering both the mean and covariance of the features, Mahalanobis distance accounts for the correlations between variables, making it well-suited for identifying deviations indicative of seizure-related patterns. This distance can be used to recognize instances that deviate substantially from the expected norm or distribution, making it applicable in various fields, including audio-based seizure detection.

In this research, we considered the features extracted from non-seizure segments, occurring prior to the seizure epoch, as the normal or reference class, and the features extracted from the seizure labelled windows as the target class to be detected.

The following equation represents the Mahalanobis distance md, defined as:

$$md_i = (x_i - \mu)^T \Sigma^{-1} (x_i - \mu) \tag{10}$$

- μ is the mean vector of the normal (or reference) class, which in this case is a reference period formed by different segments labelled as "non-seizure".

- x represents the feature vector to be tested.

- Σ is the covariance matrix of the reference (or normal) class.

In the context of anomaly detection:
- If md is small, it suggests that the observation vector x is close to the mean of the reference class, implying a normal or expected condition.
- If md is large, it indicates that the observation data is substantially different from the mean of the reference class. This could suggest the presence of a seizure or another anomalous event.

It's important to note that the choice of the reference class is a critical step in the application of Mahalanobis distance for seizure detection.

2.8.2 Machine Learning

2.8.2.1 k-Nearest Neighbours

k-Nearest Neighbours (k-NN) is a non-parametric machine learning algorithm currently used in a supervised learning context. The term “non-parametric” refers to the fact that k-NN does not make explicit assumptions about the underlying data distribution. This method makes predictions based on the local neighbourhood of data points in the feature space. In the context of supervised learning, k-NN is used to make predictions for new, unseen instances based on the majority class of their k-nearest neighbours in the training dataset. It requires labelled data during the training phase, where each data point is associated with a class label.

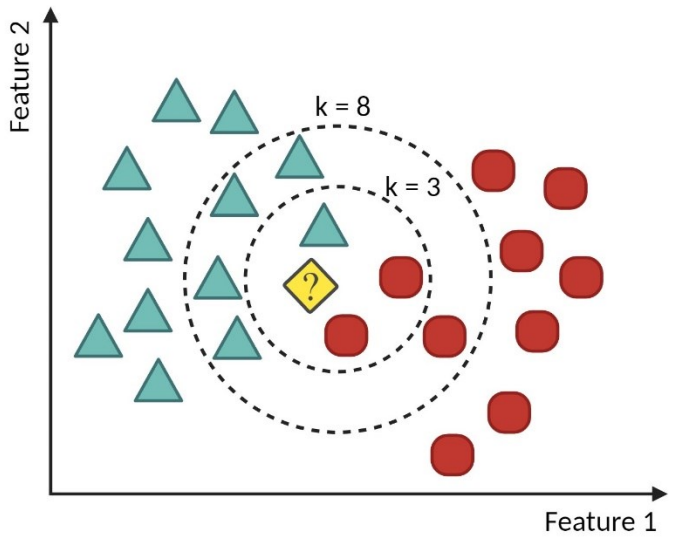

Figure 13. Visual representation of how the algorithm of k-Nearest-Neighbours works.

The anomaly score of a point is determined by taking the distance, normalized by the distances in that local neighbourhood, between the point and its k-nearest neighbours in the normal class [42].

Using the k-NN algorithm, a data point was assigned to the target class or not by obtaining a majority vote from its k most similar data points [43]. The measure of similarity used in this study is the Euclidean distance between x_{tr} (training data) and x_i (test data point or query point):

$$d(x_{tr}, x_i) = \|x_{tr} - x_i\| = [(x_{tr} - x_i)^T (x_{tr} - x_i)]^{1/2} \qquad (11)$$

A test data point x_i is considered a member of the target class if its local density is greater than or equal to that of its nearest neighbour in the training set [42], calculated with the following equation:

$$l(x_i, k) = \frac{d\big(x_i, NN(x_i, k)\big)}{d\big(NN(x_i, k), NN(NN(x_i, k), k)\big)} \qquad (12)$$

which shows that the distance from object x_i to its nearest neighbour in the training set $NN(x_i, k)$ is compared with the distance from this nearest neighbour $NN(x_i, k)$ to its nearest neighbour $NN(NN(x_i, k), k)$ [44].

$l(x_i, k)$ represents the local density of test set data point x_i. The default threshold of 1 was used to obtain the predicted labels based on the local density of the points of the test set evaluating $l(x_i, k) > 1$.

The k-NN algorithm is known for being simple to implement. However, the choice of the k value can significantly impact the algorithm's performance. Smaller k values tend to yield predictions that are more sensitive to noise in the data, while larger k values may result in smoother predictions but are less sensitive to local variations [44]. Therefore, the selection of k may be a critical aspect in the application of k-NN. Here, given the reduced available number of segments, we chose a k value of 5 as a trade-off between bias and variance.

2.8.2.1.1 Weighted k-Nearest Neighbours

In scenarios where the impact of nearby points on the prediction should be weighted according to their closeness or significance, weighted k-NN is utilized as an alternative to regular k-NN. Weighted k-NN assigns varying weights to each neighbour according to their proximity or similarity to the query point [45], instead of treating them all identically.

After selecting the k-nearest neighbours like in unweighted k-NN, the algorithm assigns a weight to each of the k-nearest neighbours inversely proportional to their distance from the query point. The weight $w(\cdot)$ is computed as follows:

$$w(x_i, NN(x_i, k)) = \frac{1}{d\big(x_i, NN(x_i, k)\big)} \tag{13}$$

In simpler terms, neighbours that are closer to the query point receive higher weights and neighbours that are farther away receive lower weights [46]. Doing so, the contributions of neighbours are weighted to determine the predicted class. The prediction is more influenced by neighbours who are closer.

In weighted k-NN, the local density of a test data point is calculated with the following formula:

$$l_w(x_i, k) = \frac{w(x_i, NN(x_i, k))\, d\big(x_i, NN(x_i, k)\big)}{w\big(NN(x_i, k), NN(NN(x_i, k), k)\big)\, d\big(NN(x_i, k), NN(NN(x_i, k), k)\big)} \tag{14}$$

Because of its subtle advantages in managing various types of data, weighted k-NN is frequently chosen over its unweighted k-NN in a variety of circumstances. The understanding that not every neighbouring point should contribute equally to the decision-making process is one of the main arguments in favour of weighted k-NN. The algorithm can better reflect the variable importance of neighbours by assigning different weights based on each neighbour's proximity. Because weighted k-NN may address imbalances by assigning greater weight to neighbours from minority classes, it is advantageous in situations where specific classes are outnumbered.

2.8.2.2 Linear Support Vector Machine

A Linear Support Vector Machine (SVM) is a powerful supervised learning algorithm used for classification and regression tasks. Its primary objective is to find the optimal hyperplane that effectively separates different classes in the feature space. The term "support vector" refers to the data points that are crucial in determining the location and orientation of the hyperplane. This margin maximization results in a robust decision boundary, making the model less sensitive to outliers.

The Linear Support Vector Machine (SVM) operates by finding the optimal hyperplane that separates different classes in the feature space. Mathematically, the objective is to define a hyperplane represented by the equation:

$$w \cdot x + b = 0 \tag{15}$$

Where w is the weight vector, x is the input feature vector, and b is the bias term. The decision function for a new data point x is then determined by the sign of $w \cdot x + b$. The decision boundary is formed by the hyperplane, and the distance from the data point to the decision boundary determines its confidence or likelihood of belonging to a particular class.

The goal of the SVM is to maximize the margin between the hyperplane and the nearest data points (support vectors) from each class. The margin is defined as the perpendicular distance from the hyperplane to the closest data point. For a linear SVM, the margin is given by: $\frac{2}{\|w\|}$ where $\|w\|$ represents the Euclidean norm of the weight vector w.

The optimization problem of the linear SVM can be formulated as follows:

$$minimize\ \frac{1}{2}\|w\|^2$$

Subject to the constraints: $y_i(w \cdot x_i + b) \geq 1$ for all training samples. Here, y_i is the class label (-1 for the negative class and 1 for the positive class), x_i is the feature vector of the i-th training example, and the constraint ensures that each data point lies on the correct side of the decision boundary with a margin of at least 1.

The Lagrange multiplier method is often used to solve the optimization problem, resulting in a set of coefficients that define the hyperplane. The support vectors are the data points that have non-zero Lagrange multipliers and are crucial in determining the position and orientation of the hyperplane.

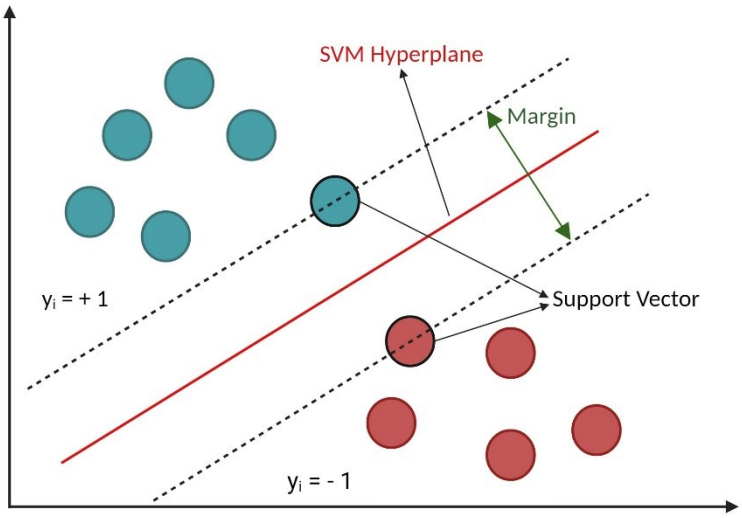

Figure 14. Visual representation of how the algorithm of linear support vector machine model works.

Although more complex kernels can be used (e.g. polynomial or Gaussian), in this study we used a linear kernel simply to avoid further tuning parameters (order of polynomial or the Gaussian's parameter) procedures.

2.8.2.3 Decision Trees

Decision trees are useful tools in seizure detection tasks because they help identify trends and provide well-informed predictions based on features that correspond to the presence or absence of seizures. The decision tree method divides the input space recursively into distinct subsets. It can be tuned with hyperparameters such as maximum number of splits and split criterion. The algorithm is guided in choosing features that efficiently distinguish between seizure and non-seizure through the Gini index, which is utilized as a measure of impurity. The tree determines which characteristic minimizes the Gini index at each node, producing branches that, depending on the features chosen, represent various conditions.

The structure of a decision tree consists of three main components: root nodes, decision nodes, and leaf nodes as illustrated in the following figure.

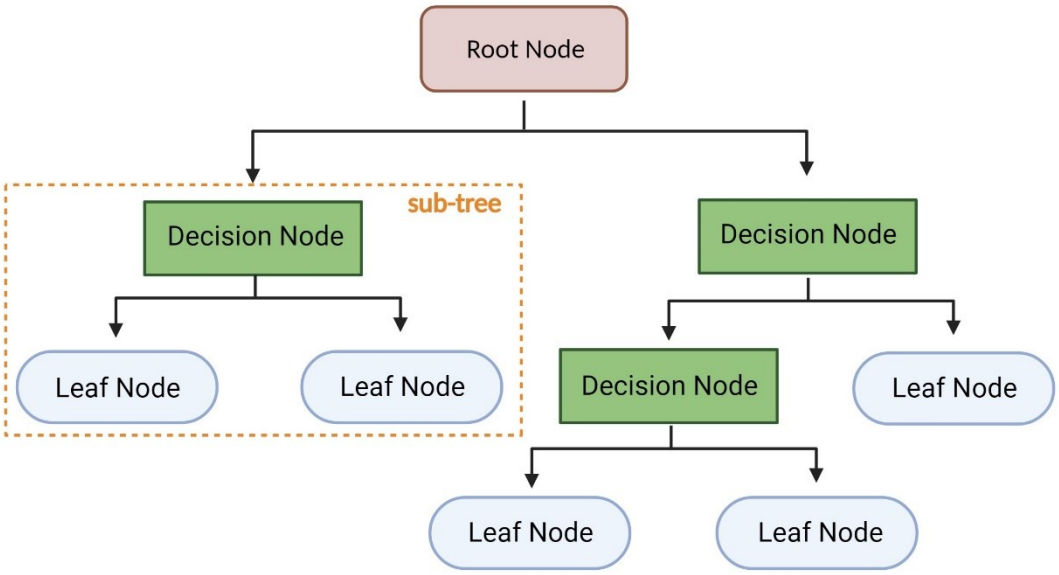

Figure 15. Visual representation of a decision tree model.

1. Root Node:

- The root node is the topmost node in the tree.
- It represents the entire dataset or a subset of it.
- The root node contains a decision based on a specific feature that best splits the data into two or more subsets.
- The feature and the corresponding split point are determined by choosing the combination that optimally separates the data according to a predefined criterion, such as Gini impurity.

2. Decision Nodes:

- Decision nodes are non-terminal nodes in the tree, occurring after the root node.
- Each decision node represents a decision based on a particular feature and its split point.
- The decision node has branches leading to child nodes, each corresponding to a possible outcome of the decision.
- The process of creating decision nodes is repeated recursively, resulting in a hierarchical structure.

Stopping conditions prevent the tree from becoming too deep and overfitting the training data. These conditions may include a maximum depth for the tree, a minimum number of samples in a node, or a threshold for impurity.

3. Leaf Nodes:

- When a stopping condition is met, the node becomes a leaf node
- Leaf nodes are terminal nodes at the bottom of the tree.
- They do not contain any decisions but rather represent the predicted outcome.

- Each leaf node is associated with a specific label.

The path from the root node to a leaf node defines a decision path or rule that can be followed to make predictions. The final prediction is the value associated with the leaf node reached.
To avoid an overfitting of the model that could result in poor accuracy, here we set a maximum number of splits equal to 4.

2.8.2.4 Neural Networks

In the pursuit of detecting epileptic seizures from audio recordings, two neural network models are employed, each featuring distinct configurations of fully connected layers. Neural networks are a class of machine learning models inspired by the structure and function of the human brain [47]. In these models, neurons are the basic computational units of a neural network and they are organized into layers.
There are typically three types of layers:

- Input layer: this is the initial layer of the neural network. It receives the raw input data or features that are fed into the network for processing. Neurons in the input layer correspond to features in the input data. and they simply feed those values to every neuron in the first hidden layer.

- Output layer: this is the final layer of the neural network. It produces the network's output based on the computations performed in the hidden layers. In detection tasks, such as distinguishing between seizure and non-seizure instances, the output layer produces predictions in the form of probabilities assigned to each class.

- Hidden layers: hidden layers are layers in a neural network that come between the input and output layers. They are called "hidden" because they are not directly observable from the network's input or output. Hidden layers enable neural networks to learn complex patterns in the data by performing nonlinear transformations on the input. In a fully connected layers (specific configuration of hidden layer that we used in this study) each neuron is connected to every neuron in the previous layer and each connection has a weight associated with it. This means that the output of each neuron in the previous layer serves as input to every neuron in the fully connected layer. Fully connected layers are common in many neural network architectures and are often used to learn complex relationships between features in the data. Each neuron in a hidden layer receives input from all neurons in the previous layer, performs a computation (weighted sum of inputs plus bias), and then applies an activation function to produce an output [35]. The hidden layers are responsible for extracting relevant features and

patterns from the input data, which are essential for making accurate predictions or classifications.

Neuron activation is defined by mathematical functions applied to the output of neurons in every layer of the structure, within a neural network. They introduce nonlinearity to the network, enabling it to learn and approximate complex relationships in the data. Activation functions are crucial because without them, neural networks would simply be linear transformations of the input data, severely limiting their ability to model nonlinear patterns [35].

ReLU, or Rectified Linear Unit, is one of the most commonly used activation functions in neural networks and we used it in this study. It is defined as:

$$f(x) = max(0, x) \tag{16}$$

In other words, ReLU takes an input x and returns the maximum of x and 0. This means that if the input is positive, ReLU will output the input value unchanged, but if the input is negative, it will output 0 [35].

ReLU is computationally efficient and easy to implement. It involves a simple comparison and selection operation, making it faster to compute compared to other activation functions.
Moreover, it introduces sparsity (a significant number of zero values) into the network by setting all negative values to zero. This sparsity can help prevent overfitting by reducing the number of active neurons and encouraging the network to focus on the most relevant features in the data.

As mentioned at the beginning of the paragraph, we employed two neural network models in this research project: one with a single fully connected layer (**Figure 16**) and one with two fully connected layers (**Figure 17**).

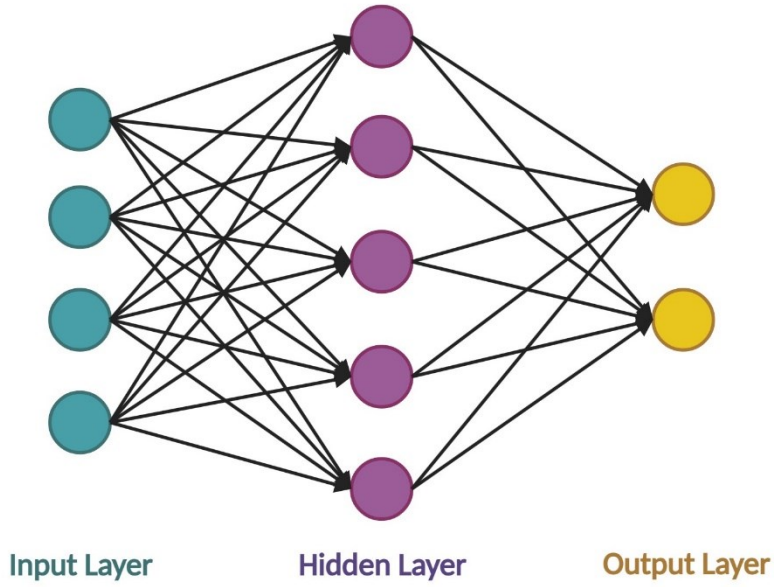

Figure 16. Visual representation of a 1-hidden layer neural network model.

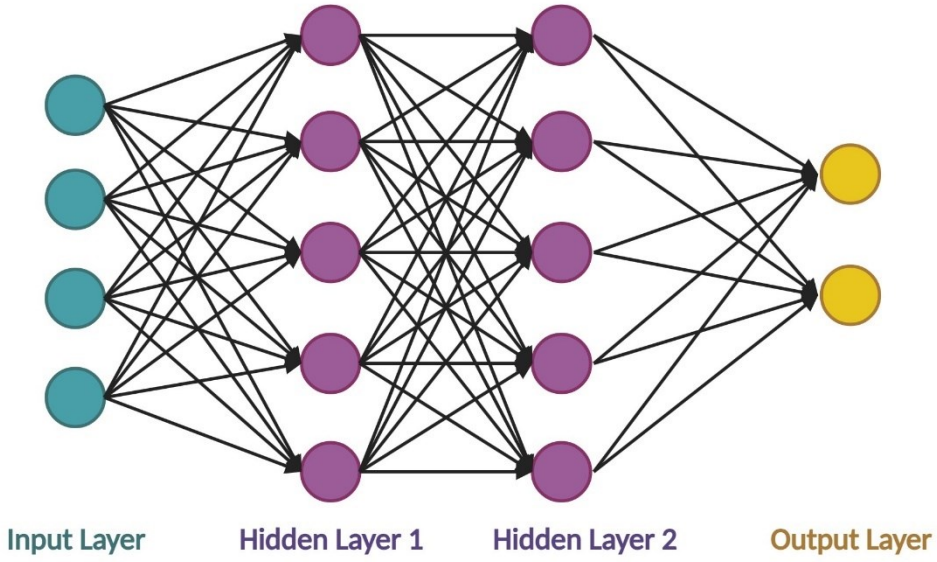

Figure 17. Visual representation of a 2-hidden layers neural network model

On the basis of the number of features fed to the neural networks models, here we used the following parameters: hidden layers size 10 (for time-domain features and for both models), hidden layers size 100 (for time-frequency domain features and for both models). For the activation function ReLu, we set the iteration limit equal to 1000 (for time-domain features patients) and of 10000 (for time-frequency domain features).

2.9 Training and testing procedures

Machine learning models go through an essential phase called training, during which they iteratively change their internal parameters to learn the statistical models from training data. There are usually multiple important steps in the procedure. A dataset is initially split into training and test sets. The former is used to train the model, while the test set is used to evaluate the model's performance on unseen data.

The model starts with random or predefined parameters and makes predictions on the training data. The discrepancy between these predictions and the actual outcomes is measured by a loss function. The model then iteratively changes its parameters to reduce this loss using optimization techniques. This iterative optimization process keeps going until the model reaches a point where making more changes only slightly improves it or until a certain number of iterations are completed. In order to make sure the trained model can generalize to new, unseen data, it is next assessed using the test set. Training procedure is crucial to produce a model that captures significant patterns in the training data and generalizes well to new data.

Cross-validation. In machine learning, cross-validation is an essential method for assessing how well predictive models work. Cross-validation, or out-of-sample testing, is a model validation technique for assessing how the results of a statistical analysis will generalize to an independent data set. This procedure generally includes sample splitting methods that use different portions of the data to test and train a model on different iterations. Here, we randomly split the data into a training set (85% of data) and test set (15% of data) and performed a cross-validation on the training set. The advantage of cross-validation over a single train-test split is that it can yield a more accurate and consistent estimate of a model's performance and it can improve its generalization to a variety of datasets.

K-fold Cross-Validation. Within the training set, further division is often conducted using k-fold cross-validation, separating the data into training and validation subsets. This approach allows for the assessment of model performance on different portions of the training data to ensure generalizability [35]. Further, cross-validation gives a better estimate of the error rate than non-cross-validated approaches at the cost of more computation [48]. Within this framework, 15% of the data is held out from the cross-validation phase as a separate test set, reserved to evaluate the model's performance on truly unseen observations, providing an independent assessment of its effectiveness.

In this study, a k-fold cross-validation with k equal to 5 was performed to validate the models.

For the k-NN method, the entire dataset, consisting of both "seizure" and "non-seizure" features, was split into training and test sets. Within the training set, a subset containing only "non-seizure" features was divided into k folds for cross-validation. During each iteration, the model was trained on the "non-seizure" data and validated on a subset containing both "seizure" and "non-seizure" features. Finally, the model was tested on the separate test set, which included samples from both classes. In this study, for all the two-classes classifiers, the training set containing features labelled both as “seizure” and “non-seizure” was divided into 5 folds to perform the cross-validation. A larger dataset per patient could have allowed larger values of k in the k-fold procedure.

2.10 Performance Evaluation

Accuracy, precision, recall, and F1 score are the four main metrics that provide a thorough evaluation of a model's performance.

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \tag{17}$$

$$Precision = \frac{TP}{(TP + FP)} \tag{18}$$

$$Recall = \frac{TP}{(TP + FN)} \tag{19}$$

$$F1\ Score = 2\frac{Precision * Recall}{(Precision + Recall)} \tag{20}$$

where TP denotes the true positives test samples, TN true negatives observations, FP false positives predictions and FN false negatives predictions.

Accuracy is a fundamental metric that assesses the overall performances of a model. It calculates the ratio of correctly classified instances (both true positives and true negatives) to the total number of instances. It is valuable for balanced datasets but may be less informative in imbalanced scenarios. Despite its wide usage, accuracy alone might be insufficient in situations where the cost of misclassifying one class is substantially higher or more critical than others.

Recall (or True Positive Rate or Sensitivity) measures the model's ability to correctly identify all positive instances relative to the total actual positive instances. It indicates how well the model can capture all positive instances and is crucial in situations where avoiding false negatives is fundamental. Sensitivity is particularly important in situations where accurate

detection of positive instances is a priority, like in the present study, where we aim at detecting the seizure events.

Precision is a key metric that evaluates the accuracy of positive predictions made by a model. It is calculated as the ratio of true positive instances to the sum of true positives and false positives. It is particularly relevant in scenarios where minimizing false positives is crucial. High precision indicates a low rate of false positives, making it essential in applications where the cost or impact of misclassifying positive instances is significant.

The F1 score is a composite metric that balances precision and recall, offering a comprehensive evaluation of a model's performance. It is the harmonic mean of precision and recall, providing a single value that considers both false positives and false negatives. F1 score is valuable in situations where achieving a balance between precision and recall is essential. It is particularly useful in imbalanced datasets or when there is an uneven cost associated with false positives and false negatives. All performance metrics, discussed above, range from 0 to 1, with 0 representing the worst possible performance (no correct predictions or true positives) and 1 representing the best possible performance (perfect predictions or true positives).

A common method for assessing detectors' performance over a variety of trade-offs between true positive and false positive error rates is the Receiver Operating Characteristic (ROC) curve. An established performance statistic for a ROC curve is the Area Under the Curve (AUC) [49]. The X-axis of a ROC curve shows $\%FP = \frac{FP}{(TN+FP)}$, while the Y-axis shows $\%TP = \frac{TP}{(TP+FN)}$ and it represents the family of optimal decision boundaries for the relative costs of TP and FP. The ideal point in a ROC (Receiver Operating Characteristic) curve represents the optimal point that corresponds to the maximum possible distance between the ROC curve and the diagonal reference line. This diagonal line represents the scenario of random classification, where the true positive rate is equal to the false positive rate. At the ideal point, the true positive rate (TPR or Recall) is 1, and the false positive rate (FPR) is 0. In other words, the ideal point represents a situation where the model can correctly classify all positive instances without generating any false positives. However, it's important to note that in many cases, the ideal point may not be achievable in practice as classification models inevitably make some errors. Therefore, the ROC curve is used to assess the model's performance by comparing its ability to discriminate between positive and negative classes against random classification rather than directly aiming for the ideal point.

In conclusion, performance metrics are essential for assessing how well machine learning models work since they reveal information about how well they can do different types of

classification tasks. When combined, these metrics provide a more complex picture of a model's advantages and disadvantages, which helps developers optimize and fine-tune their machine learning algorithms for certain application domains.

2.11 Effect of imbalanced data

The term imbalanced data is associated with datasets that show a dataset with a significant skewed class proportion of data [50]. As we can see from [51], [52], models generally produce a highly unequal degree of accuracy, with the majority class having near-perfect accuracy and the minority class having accuracies of 0–10 per cent. This type of outcome generates important consequences, especially in the medical industry, where imbalanced datasets are particularly common [53]. With an important disproportion of samples, the training model will spend most of its time on the majority examples and not learn enough from samples of the minority class. In the case of seizure detection, it can signify that seizure epochs (a minority class) could potentially been missed, therefore not alarming caregivers.

Therefore, a system that can provide high accuracy for the minority class without compromising the accuracy of the majority class is needed for this particular setting. In cases of imbalanced datasets, the traditional evaluation technique based on a single assessment criterion, such as overall accuracy for instance, does not provide enough information [50]. Therefore, for definitive assessments of performance in the presence of imbalanced data, more informative metrics are required, such as the receiver operating characteristics curves, recall, precision [50] and F1 score.

Nevertheless, some studies have demonstrated that despite the limited instances in the minority class compared to the majority class, some models are capable of accurately learning the representation of the minority class [54], [55], [56]. These results suggest that learning can be affected by more than just the degree of imbalance. As it happens, the main factor influencing detection deterioration is the dataset's complexity, which is amplified by the inclusion of a relative imbalance [50].

Approaches to solve class imbalance have already been applied in the Biomedical Sciences [53]. These methods apply techniques to the training data to adjust the number of samples either of the majority class or the minority class, hence improving the distribution of classes. Currently, the data-level approach primarily focuses on balancing training data across classes in the data space by resampling.

Under-sampling and over-sampling are the two main categories into which resampling methods can be divided [53]. When using under-sampling techniques, samples from the majority class

are eliminated until each class has roughly the same amount of samples [57]. Nevertheless, it is unavoidable that some samples that are significant to the training model may be missed when the dataset is under-sampled. Some under-sampling methods are Random Under-Sampling (RUS), Cluster Centroid and Edited Nearest Neighbours (ENN).

To achieve an equal class distribution while reinforcing class boundaries, over-sampling techniques create additional samples based on samples from the minority class [58]. Oversampling, however, can result in overfitting as it synthesizes or replicates a smaller amount of data [59]. Also, the training time increases together with the increase of samples. Over-sampling methods include, for instance, Random Over-Sampling (ROS), Synthetic Minority Over-Sampling Technique (SMOTE) and Adaptive Synthetic (ADASYN) algorithms [53].

Sampling techniques are not the only way to handle class imbalance. Machine learning models based on single (majoritarian) classes, for example, have drawn a lot of interest. More specifically, as opposed to differentiating between instances of both positive and negative classes as in the conventional machine learning approaches, this kind of approach seeks to recognize instances by using only a single class for training [50].

2.11.1 Synthetic Minority Over-Sampling Technique (SMOTE)

In this study, we decided to implement the over-sampling technique called Synthetic Minority Over-Sampling Technique (SMOTE). It is a popular method used to address class imbalance in machine learning datasets and, as previously mentioned, it works by generating synthetic examples of the minority class.

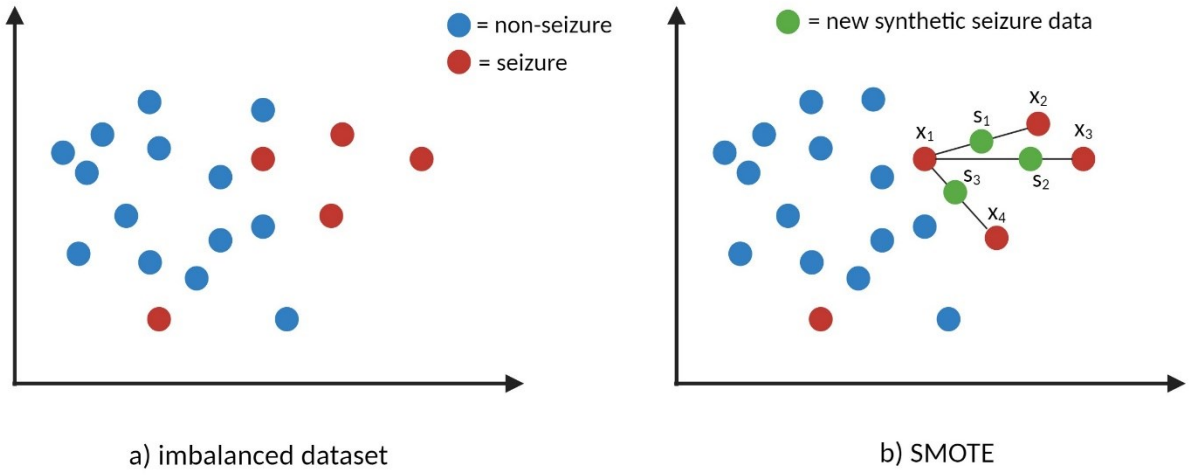

Figure 18. Visualisation of Synthetic Minority Over-Sampling Technique (SMOTE), a) imbalanced dataset priore to data augmentation and b) dataset after SMOTE. Blue circles represent non-seizure data, red circles seizure data and green circles new synthetic seizure data.

The first step is to identify the minority class within the dataset, which in our case is the class labelled as "seizure". Afterwards, the SMOTE algorithm selects a random observation from the minority class and it finds its k-nearest neighbours in the feature space, which are other observations from the same class. Then, the algorithm generates synthetic samples by linearly interpolating between a randomly selected sample and its randomly selected neighbour using a random weight. This process is repeated until the desired balance between classes is achieved. In our case, the goal was to have the minority class account for 50% of the dataset. After the split of dataset in training and test sets, we applied the SMOTE algorithm to re-balance the classes. Although k could be a hyperparameter to optimize using cross-validation, we notice that SMOTE relies on the k nearest neighbours' algorithm, which does not scale well. Optimizing this value could be therefore quite time-consuming and computationally intensive. Here, to perform a fast random sampling of the minority class, we set k =3.

By creating synthetic examples, SMOTE helps to handle the class imbalance problem without duplicating existing samples. This can prevent overfitting and improve the generalization of models' performance, especially when the minority class is underrepresented in the dataset.

CHAPTER 3: EXPERIMENTAL RESULTS

3.1 Features Visualization

As a preliminary analysis of the data, we compare the distribution of different features extracted from the audio recordings according to the labelled sound. **Figure 19** shows the box plot of time-domain features such as mean, zero crossing rate and root mean square energy derived from one patient and distributed among different sound categories. It is possible to notice that features extracted from segments of audios labelled as "ictal period" or "seizure" have higher values compared to other labels.

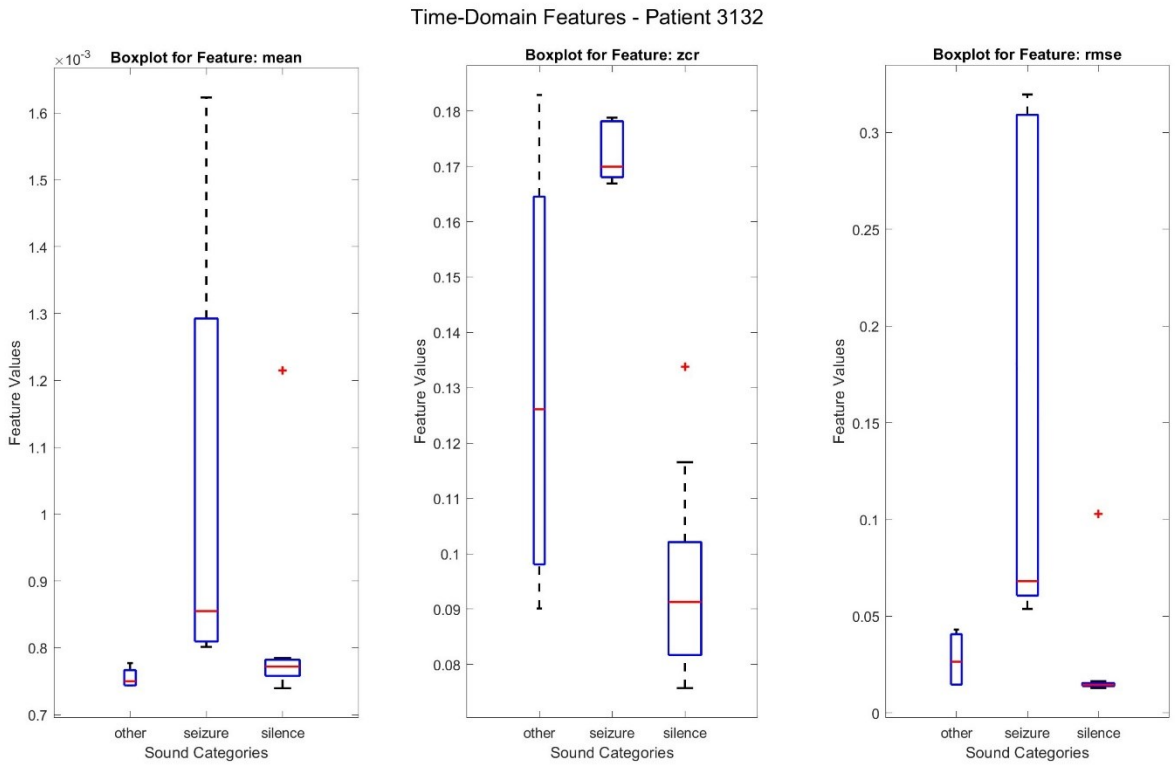

Figure 19. Boxplot of time-domain features (mean, zero crossing rate and root mean square energy) obtained from audio recordings of patient 3132 divided among its different sound labels: seizure (ictal epoch), silence (minimal sound activity) and other (undefined sounds).

Despite the clear distinction observed for this patient, **Figure 20** enables a visual summary of the distribution of the feature zero crossing rate among all patients, and all sound categories. The label "seizure" gets lost among the other labels, giving an insight into the large variability among patients.

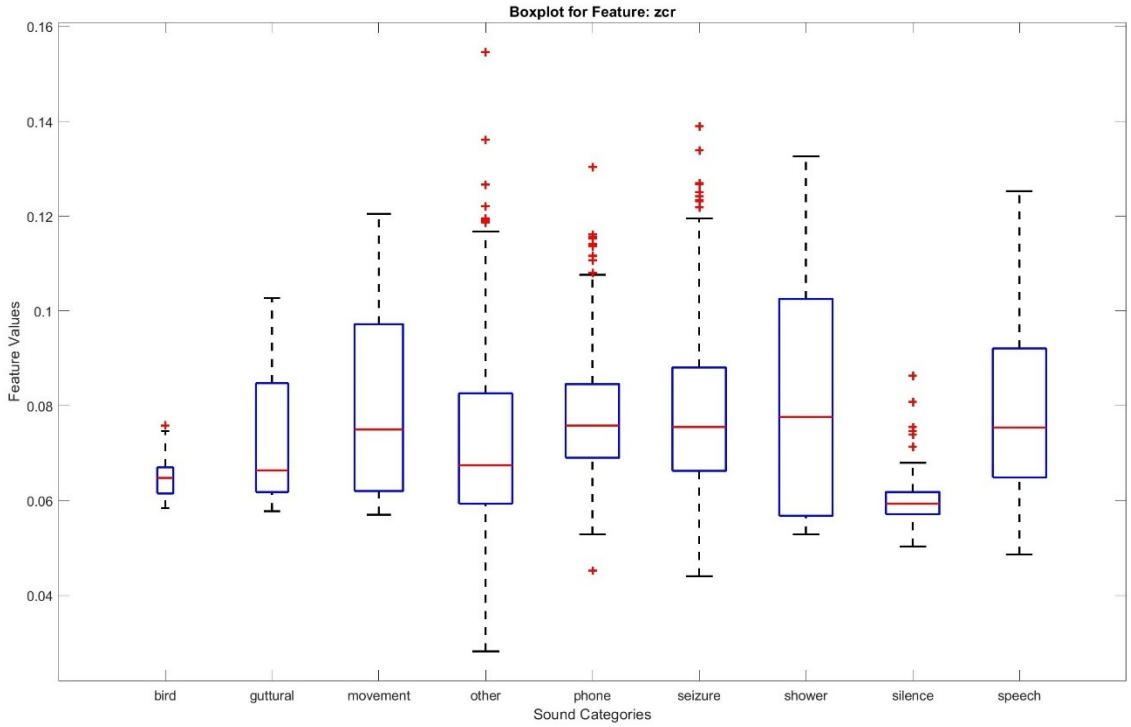

Figure 20. Boxplot of feature zero crossing rate obtained from all patients and distributed among all sound labels present in all audio recordings.

3.2 Results of Mahalanobis Distance

For illustration purposes, the Mahalanobis distance computed from recordings of one patient is shown in Figure 21. A direct comparison of the audio recordings extracted from videos with and without epileptic ictal epochs ("seizures") enables to qualitatively recognize statistical differences between the two cases. As mentioned in Chapter 2: Materials and Methods, the normal or reference class is made of segments labelled as "non-seizure".

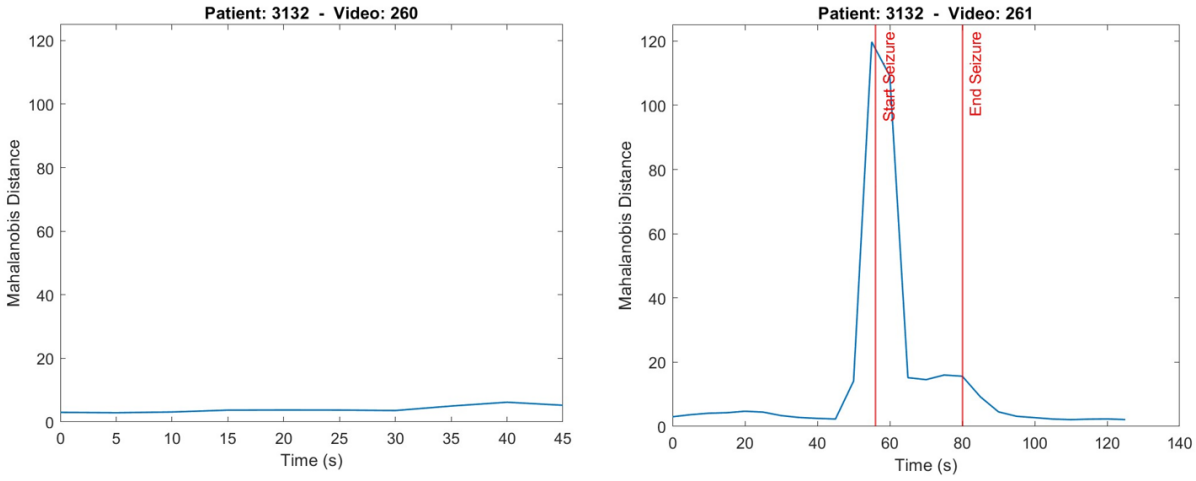

Figure 21. Mahalanobis distance of audio recordings extracted from videos 260 and 261 of patient 3132 with seizure start and end markers in red.

In video 261 of patient 3132, we observe an increase in the Mahalanobis distance value, indicating instances that deviate from the expected or reference (interictal, seizure-free) distribution. This peak aligns with the red marker at the onset of the seizure. Subsequently, we notice a gradual decrease in the Mahalanobis distance towards the end of the seizure. Quantitatively, the Area Under the Curve (AUC) for patient 3132 was assessed to be 0.98.

3.3 Performance Metrics on the original imbalanced data

3.3.1 Performance metrics for Patient 3132 (imbalanced data)

The algorithm used to implement the k-NN model provided the following results.

Figure 22 shows the local density of datapoints for patient number 3132 along with a region highlighted in red on the graph where the seizure episode occurs. This highlighting is based on the labels associated with the data points, offering a visual representation of the seizure event within the dataset.

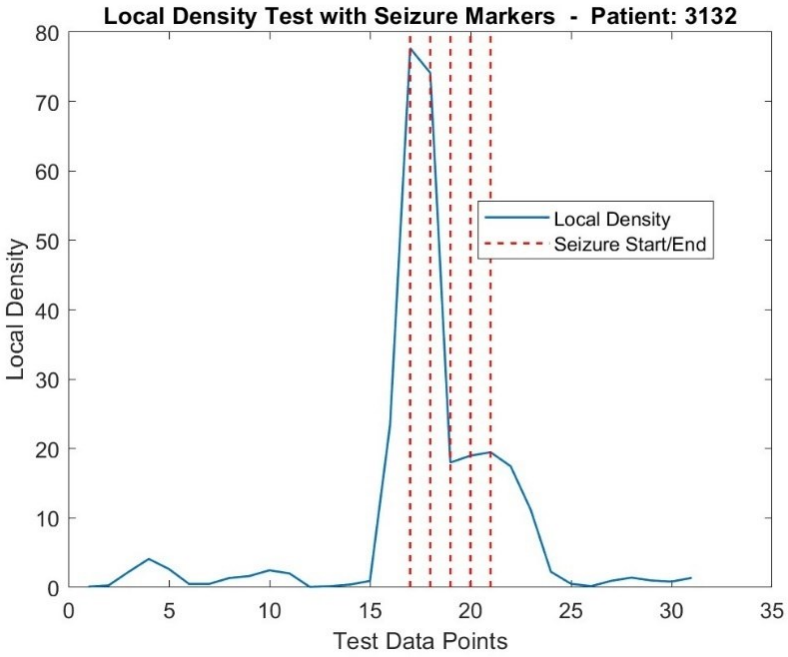

Figure 22. Local density of datapoints of patient 3132 obtained with k-NN model (k=5) in red the seizure start and end markers.

The data were predicted based on a specified threshold (as mentioned in Chapter 2: materials and methods): a label of 0 indicated membership in the normal class if the local density was below the threshold, while a label of 1 was assigned for an outlier status if the local density exceeded the threshold. Afterwards, a confusion matrix of the test set (**Figure 23**) was

calculated by comparing the predicted labels with the real labels, where a label of 0 denoted features extracted from windows labelled as “non-seizure”, and a label of 1 denoted features extracted from windows labelled as “seizure”.

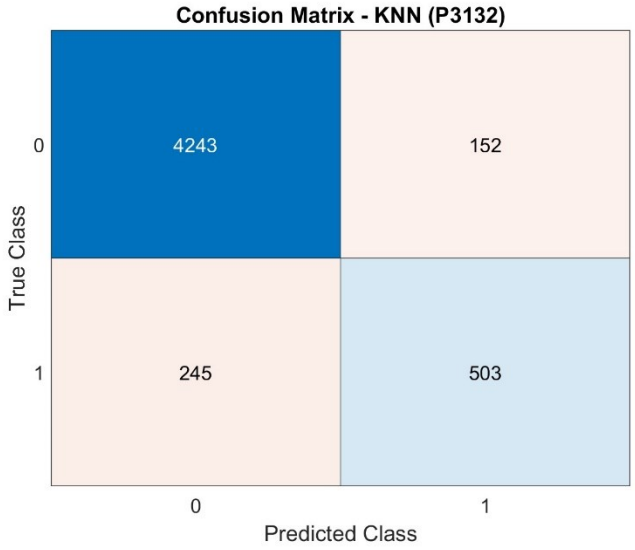

Figure 23. Confusion matrix of k-NN model generated from the test set (15%) of patient 3132. The model utilized time-frequency domain features as input, where each signal window yielded a feature vector. The size of the feature vector depended on the number of frames (partitions) and the degree of overlap used to estimate the Mel spectrogram and the related parameters.

Upon completing the k-fold cross-validation of the training set, the confusion matrix of the test set was used to derive the classifier metrics of accuracy, precision, recall, and F1 score. Performance metrics for the patient 3132 are reported in **Table 5**.

	k-NN			
P3132	**Accuracy**	**Precision**	**Recall**	**F1 Score**
Time-domain Features	0.94	0.84	0.69	0.77
Time-Frequency domain Features	0.92	0.77	0.67	0.72

Table 5. Accuracy, Precision, Recall and F1 Score obtained for k-NN model and patient 3132 both using time-domain features and time-frequency domain features (original imbalanced dataset).

The accuracy of the k-NN model for patient 3132 is considerably high, with values of 0.94 and 0.92 for time-domain features and time-frequency domain features, respectively. This indicates that the model correctly classified a large proportion of instances as either seizure or non-seizure events for this specific patient. The precision values are 0.84 and 0.77 and recall values are 0.69 and 0.67 for time-domain features and time-frequency domain features respectively. This

difference between precision and recall illustrates the model's tendency to avoid false positive predictions over capturing all positive instances in the dataset. The F1 scores are 0.77 and 0.72 for the two sets of features, indicating robust performance in terms of both precision and recall. Comparing the results between time-domain features and time-frequency domain features, we observe that the values of the performance metrics remain relatively consistent. However, it is worth noting that slightly higher results are observed when utilizing time-domain features as input.

Figure 24 represents a scatterplot of predictions made by the weighted k-NN model on the training set when utilizing time-domain features as an input.

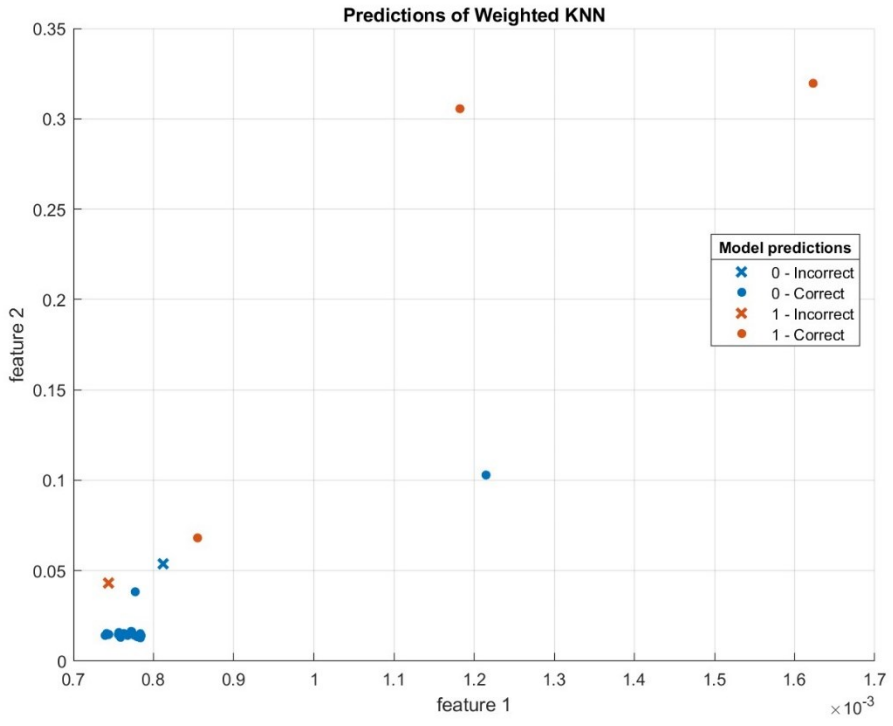

Figure 24. Scatter plot of predictions of model Weighted k-Nearest Neighbours implemented with time-domain features of patient 3132 as input. Blue points indicate observations from the "non-seizure" class, while red points indicate the "seizure" class.

This plot typically illustrates the relationship between the predicted values generated by the model and the actual values of the target variable. In the context where feature 1 (for example the mean) and feature 2 (for example the standard deviation) are plotted on the x-axis and y-axis respectively, the scatter plot represents how well the model's predictions align with the true labels across these two features. Each point on the scatter plot represents an observation in the training set. The position of the point is determined by the values of feature 1 (x-axis) and

feature 2 (y-axis) for that observation. Points correctly predicted as belonging to the “non-seizure” class are represented by the colour blue and the dot shape, while red dots are the datapoints correctly predicted as belonging to the “seizure” class. Whenever a point is incorrectly predicted it is represented by a x shape. Ideally, the points should cluster closely around a diagonal line, indicating that the model's predictions closely match the true values. The observed dispersion or misalignment out of the diagonal suggest that the model's predictions are less accurate.

For illustration purposes of the classification obtained with the different models, in **Figure 25**, the confusion matrices depict the performance of weighted k-NN, Linear SVM, one-layer Neural Network, bi-layer Neural Network, and decision tree. These matrices are obtained utilizing the test set of the time-frequency domain features dataset of patient 3132. They are presented as they provide a more indicative evaluation of the models' performance compared to those obtained using time-domain features, which are instead quite limited in terms of number of observations of the test set.

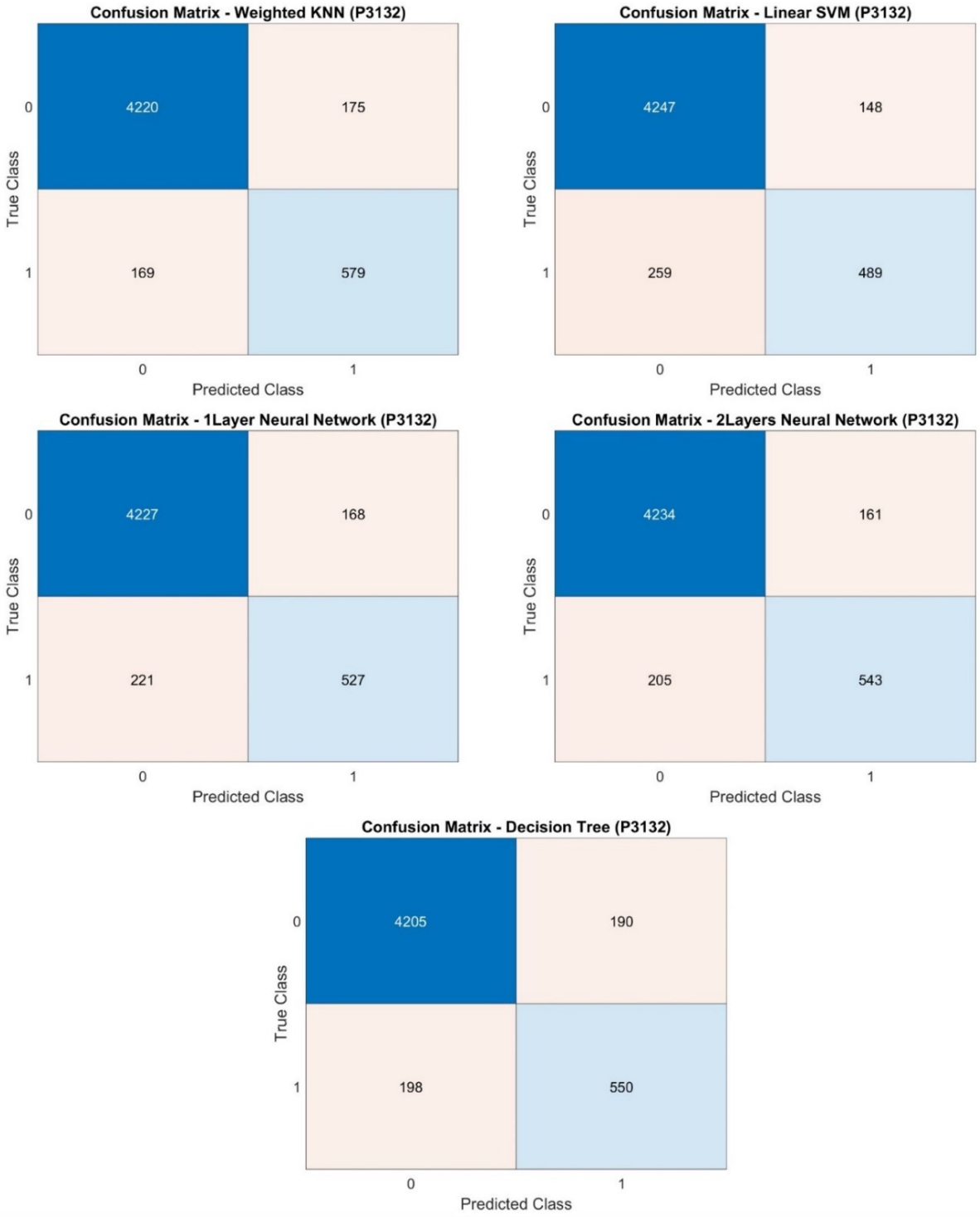

Figure 25. Confusion matrices of weighted k-NN, linear SVM, one-layer NN, bi-layer NN and decision tree models generated from the test set of patient 3132 utilizing time-frequency domain features as input (original imbalanced dataset).

The performance metrics derived from the respective confusion matrices are summarized in Table 6 and Table 7, distinguishing between input features from the time domain and those from the time-frequency domain.

P3132	Models	Accuracy (k-fold cross validation)	Accuracy (Test)
Time-domain Features	Weighted kNN	0.94	1
	Linear SVM	0.87	0.80
	Decision Tree	0.94	1
	1-Layer NN	0.90	1
	2-Layers NN	0.90	1
Time-Frequency domain Features	Weighted kNN	0.94	0.93
	Linear SVM	0.92	0.92
	Decision Tree	0.93	0.92
	1-Layer NN	0.93	0.92
	2-Layers NN	0.93	0.93

Table 6. Accuracy obtained after training and test of models weighted k-NN, linear SVM, decision tree, 1-Layer NN and 2-Layers NN performed with time-domain features and time-frequency features as input for patient 3132 (original imbalanced dataset).

The weighted k-NN model achieved the highest accuracy across both feature sets, indicating its ability to correctly classify seizure and non-seizure instances. Although all models reached very high accuracy values, Linear SVM obtained a lower accuracy when using time-domain features, specifically 0.80 in value.

Despite the higher values obtained for this patient, it's essential to consider whether accuracy is an appropriate metric, especially when dealing with imbalanced datasets, as it may not accurately reflect the model's performance. Precision, recall, F1 score and AUC (**Table 7**) may provide additional insights into the performance of the models. We notice, however, that the reduced number of test samples rendered impossible to estimate the precision and F1score for this patient. Specifically, linear SVM model exhibited great challenges, particularly in the time-domain feature set. In fact, the number of true positives reported is equal to 0, making it impossible to compute precision and F1 scores for the SVM models.

The AUC values of 1 achieved by the weighted k-NN, decision tree, and neural network models for time-domain features indicate perfect discriminatory power. This suggests that these models were able to perfectly separate seizure and non-seizure instances for patient 3132 based on time-domain features alone. The linear SVM and neural network models also achieved high AUC values (0.98 and 1, respectively) for time-domain features. Despite the bias induced by the imbalanced data, reported values highlight the suitability of some of these models for seizure detection based on time-domain features. The AUC values obtained for models using time-

frequency domain features are lower compared to those obtained using time-domain features. However, they still range from 0.88 to 0.92, indicating great discriminatory power.

P3132	Models	Precision	Recall	F1 Score	AUC
Time-domain Features	Weighted kNN	1	1	1	1
	Linear SVM	/	0	/	0.98
	Decision Tree	1	1	1	1
	1-Layer NN	1	1	1	1
	2-Layers NN	1	1	1	1
Time-Frequency domain Features	Weighted kNN	0.77	0.77	0.77	0.88
	Linear SVM	0.77	0.65	0.71	0.90
	Decision Tree	0.74	0.74	0.74	0.90
	1-Layer NN	0.76	0.70	0.73	0.92
	2-Layers NN	0.76	0.70	0.73	0.92

Table 7. Precision, Recall, F1 Score and Area Under the Curve (AUC) obtained after test of models weighted k-NN, linear SVM, decision tree, 1-Layer NN and 2-Layers NN performed with time-domain features and time-frequency features as input (original imbalanced dataset).

The F1 score provides a balance between precision and recall, considering both false positives and false negatives and it is particularly useful when the dataset is imbalanced, as the data tested here. The weighted k-NN model achieved the overall highest F1 scores with time-domain features and 0.77 with time-frequency domain features, indicating good performance. Decision tree and neural network models, including single and double hidden layer architectures, showed consistent performance across both feature sets. They achieved good precision, recall, and F1 scores between 0.70 and 0.76, indicating their potential for seizure detection from audio recordings.

3.3.2 Performance metrics averaged among all patients (imbalanced data)

To have an overall assessment of the models' performance on the original *imbalanced* dataset, **Table 8** displays the performance metrics obtained by averaging among all patients.

All Patients	Models	Accuracy	Precision	Recall	F1 Score	AUC
Time-domain Features	Weighted kNN	0.97	0.83	0.70	0.75	0.93
	Linear SVM	0.87	/	0	/	0.93
	Decision Tree	0.95	0.83	0.74	0.79	0.94
	1-Layer NN	0.92	0.70	0.82	0.76	0.95
	2-Layers NN	0.88	0.68	0.82	0.75	0.91
Time-Frequency domain Features	Weighted kNN	0.93	0.80	0.79	0.80	0.92
	Linear SVM	0.90	0.77	0.65	0.71	0.86
	Decision Tree	0.88	0.75	0.58	0.67	0.83
	1-Layer NN	0.88	0.78	0.68	0.73	0.83
	2-Layers NN	0.88	0.78	0.68	0.73	0.84

Table 8. Mean values of Accuracy, Precision, Recall, F1 Score and AUC averaged among all patients with time-domain features and time-frequency domain features as input of weighted k-NN, linear SVM, decision tree, 1-layer NN and 2-layers NN (original imbalanced dataset).

Table 9 presents the minimum and maximum values of the performance metrics acquired from all patients. As previously mentioned, some of these high performances are certainly biased by the imbalance nature of the dataset. In the next section, we will address this issue.

All Patients	Models	Accuracy	Precision	Recall	F1	AUC
Time-domain Features	Weigthed kNN	[0.92 1]	[0.33 1]	[0.15 1]	[0.20 1]	[0.82 1]
	Linear SVM	[0.75 0.97]	/	0	/	[0.85 0.98]
	Decision Tree	[0.83 1]	[0.33 1]	[0.31 1]	[0.31 1]	[0.83 1]
	1-Layer NN	[0.75 1]	[0.33 1]	[0.29 1]	[0.31 1]	[0.83 1]
	2-Laers NN	[0.58 1]	[0.33 1]	[0.31 1]	[0.31 1]	[0.72 1]
Time-Frequency domain Features	Weigthed kNN	[0.88 0.99]	[0.77 0.90]	[0.77 0.86]	[0.77 0.88]	[0.88 0.95]
	Linear SVM	[0.78 0.98]	[0.77 0.77]	[0.65 0.65]	[0.71 0.71]	[0.72 0.90]
	Decision Tree	[0.72 0.98]	[0.74 0.78]	[0.10 0.74]	[0.44 0.74]	[0.57 0.93]
	1-Layer NN	[0.72 0.98]	[0.76 0.85]	[0.60 0.70]	[0.73 0.73]	[0.57 0.92]
	2-Laers NN	[0.72 0.98]	[0.76 0.85]	[0.60 0.70]	[0.73 0.73]	[0.59 0.92]

Table 9. Minimum and maximum values of Accuracy, Precision, Recall, F1 Score and AUC registered among all patients with time-domain features and time-frequency domain features as input of weighted k-NN, linear SVM, decision tree, 1-layer NN and 2-layers NN (original imbalanced dataset).

3.4 Performance Metrics on balanced data (SMOTE algorithm)

To deal with the imbalance we have applied the oversampling algorithm of SMOTE previously described. After implementing the SMOTE algorithm to balance the dataset, we proceeded to train and test all machine learning models using both sets of features.

3.4.1 Performance metrics for Patient 3132 (balanced data)

To illustrate the effect of imbalance on the classification, Figure 26 shows the confusion matrices for all the tested models, with time-frequency domain features estimated from recordings of patient 3132.

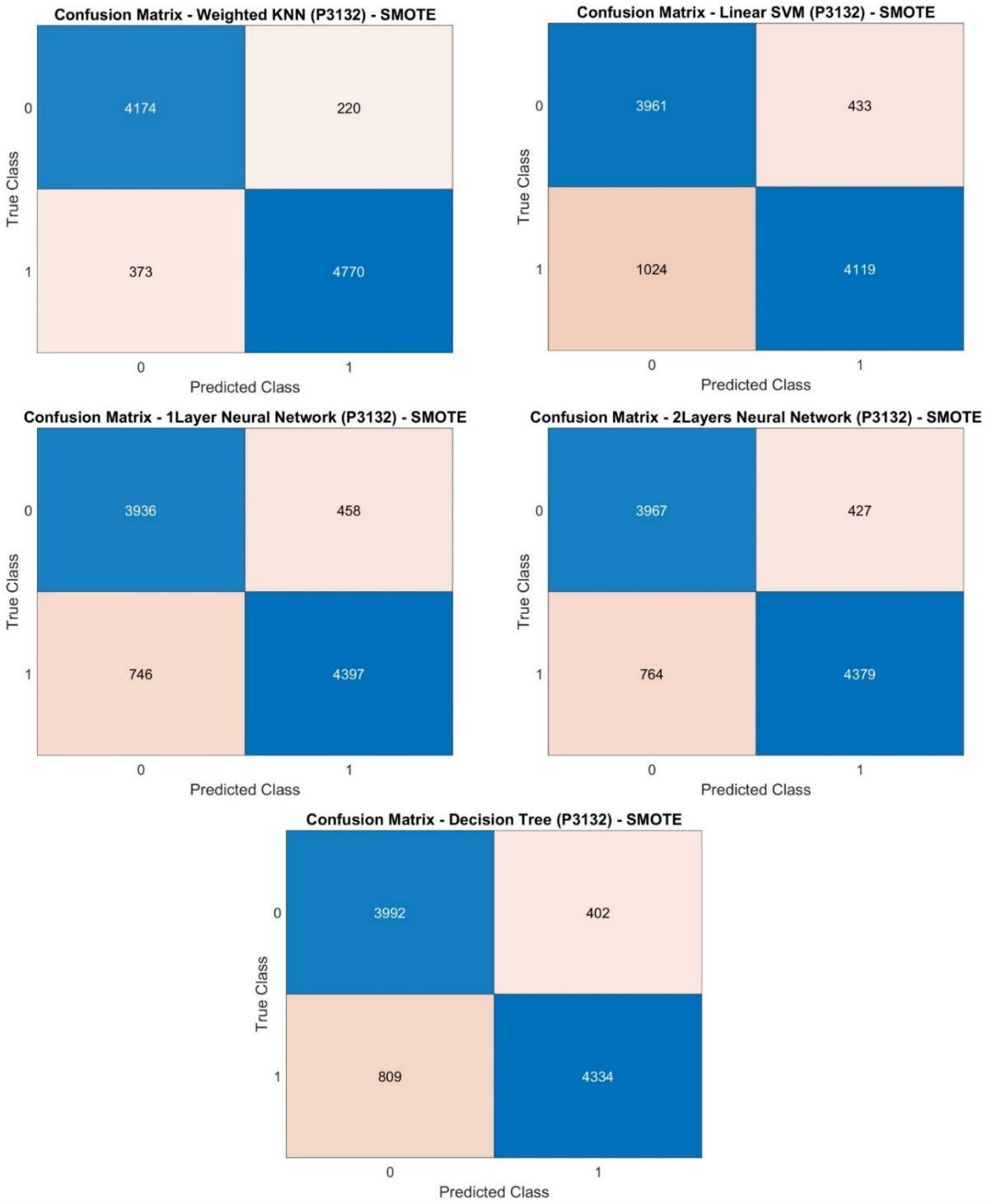

Figure 26. Confusion matrices of weighted k-NN, linear SVM, one-layer NN, bi-layer NN and decision tree models generated from the test set of patient 3132 utilizing time-frequency domain features as input (balanced dataset).

As seen in Table 10, across all machine learning models, consistent good performances were observed for patient 3132 when utilizing time-domain features. The metrics revealed an accuracy of 0.90, a precision of 0.83, a recall of 1, an F1 score of 0.90, and an AUC of 0.90. When employing frequency and time-frequency domain features, the models obtained less consistent performance. We can compare these values with those, certainly overestimated, obtained from imbalance data and reported in Table 7. We notice, however, that the new results remained within an acceptable range between 0.83 and 0.96 for all performance metrics.

P3132	Models	Accuracy	Precision	Recall	F1 Score	AUC
Time-domain Features	kNN	0.90	0.83	1	0.90	0.90
	Weighted kNN	0.90	0.83	1	0.90	0.90
	Linear SVM	0.90	0.83	1	0.90	0.90
	Decision Tree	0.90	0.83	1	0.90	0.90
	1-Layer NN	0.90	0.83	1	0.90	0.90
	2-Layers NN	0.90	0.83	1	0.90	0.90
Time-Frequency domain Features	kNN	0.90	0.90	0.92	0.91	0.96
	Weighted kNN	0.94	0.96	0.93	0.94	0.96
	Linear SVM	0.85	0.9	0.8	0.85	0.89
	Decision Tree	0.87	0.92	0.84	0.88	0.92
	1-Layer NN	0.87	0.91	0.86	0.88	0.93
	2-Layers NN	0.88	0.91	0.85	0.88	0.93

Table 10. Values of Accuracy, Precision, Recall, F1 Score and AUC obtained for patient 3132 with time-domain features and time-frequency domain features as input of k-NN, weighted k-NN, linear SVM, decision tree, 1-layer NN and 2-layers NN, after balancing the dataset implementing the SMOTE algorithm.

3.4.2 Performance metrics averaged among all patients (balanced data)

When averaging among all patients (Table 11), the values of the performance metrics are still consistent and high, except for the model linear SVM that gave slightly lower results compared to the other models.

All Patients	Models	Accuracy	Precision	Recall	F1 Score	AUC
Time-domain Features	kNN	0.89	0.87	0.95	0.90	0.92
	Weighted kNN	0.86	0.84	0.92	0.87	0.93
	Linear SVM	0.74	0.75	0.73	0.72	0.78
	Decision Tree	0.84	0.83	0.88	0.85	0.87
	1-Layer NN	0.83	0.84	0.85	0.84	0.87
	2-Layers NN	0.83	0.82	0.89	0.84	0.86
Time-Frequency domain Features	kNN	0.91	0.85	0.86	0.86	0.92
	Weighted kNN	0.96	0.96	0.90	0.93	0.97
	Linear SVM	0.86	0.79	0.74	0.76	0.85
	Decision Tree	0.87	0.82	0.81	0.79	0.86
	1-Layer NN	0.87	0.82	0.80	0.80	0.88
	2-Layers NN	0.87	0.82	0.81	0.80	0.89

Table 11. Values of Accuracy, Precision, Recall, F1 Score and AUC obtained averaged among all patients with time-domain features and time-frequency domain features as input of k-NN, weighted k-NN, linear SVM, decision tree, 1-layer NN and 2-layers NN, after balancing the dataset implementing the SMOTE algorithm.

Table 12 presents the minimum and maximum values of the performance metrics acquired from all patients.

All Patients	Models	Accuracy	Precision	Recall	F1 Score	AUC
Time-domain Features	kNN	[0.81 0.90]	[0.77 1]	[0.91 1]	[0.84 0.96]	[0.88 0.96]
	Weighted kNN	[0.76 0.93]	[0.80 0.89]	[0.77 1]	[0.80 0.92]	[0.92 0.98]
	Linear SVM	[0.55 0.90]	[0.61 0.91]	[0.35 1]	[0.44 0.90]	[0.64 0.90]
	Decision Tree	[0.76 0.92]	[0.77 0.90]	[0.77 1]	[0.78 0.92]	[0.80 0.95]
	1-Layer NN	[0.70 0.90]	[0.71 0.92]	[0.74 1]	[0.72 0.90]	[0.80 0.93]
	2-Layers NN	[0.76 0.90]	[0.76 0.88]	[0.83 1]	[0.79 0.90]	[0.74 0.95]
Time-Frequency domain Features	kNN	[0.75 0.99]	[0.75 0.90]	[0.83 0.92]	[0.79 0.91]	[0.79 0.97]
	Weighted kNN	[0.90 1]	[0.93 0.98]	[0.88 0.93]	[0.91 0.94]	[0.96 0.98]
	Linear SVM	[0.80 0.89]	[0.69 0.90]	[0.67 0.80]	[0.72 0.85]	[0.71 0.90]
	Decision Tree	[0.61 0.99]	[0.60 0.92]	[0.70 0.99]	[0.75 0.88]	[0.62 0.94]
	1-Layer NN	[0.61 0.99]	[0.60 0.91]	[0.70 0.95]	[0.74 0.88]	[0.62 0.99]
	2-Layers NN	[0.61 0.99]	[0.60 0.91]	[0.73 0.93]	[0.73 0.88]	[0.63 0.99]

Table 12. Minimum and maximum values of Accuracy, Precision, Recall, F1 Score and AUC registered among all patients with time-domain features and time-frequency domain features as input of weighted k-NN, linear SVM, decision tree, 1-layer NN and 2-layers NN (balanced dataset).

CHAPTER 4: DISCUSSION

With the number of refractory patients, including those monitored at home, and the incessant requests from patients and caregivers, a robust seizure-detection system is required. Therefore, this study aimed to assess the performance of machine learning models in detecting epileptic seizures from audio recordings, utilizing both time-domain features and time-frequency domain features. Particularly, this work presented an epileptic seizure detection method based on audio recordings of patients. In this study we assessed the performances of different statistical models including the Mahalanobis distance, k-nearest neighbours, decision tree, linear support vector machine and neural networks.

The proposed approach was tested on a dataset of audio recordings showcasing a variety of sounds recorded at the patients' rooms. This approach was chosen as the visualizations of the audio features revealed great amount of variability among patients, making a patient-specific analysis more appropriate.

Through this research, results indicated that the Mahalanobis distance is statistically different when a seizure occurs and that all models, except for the linear support vector machine, can properly classify seizures in audio data from our dataset. Performance parameters, including accuracy, precision, recall, F1 score and AUC were calculated for each patient to assess the models' effectiveness. Prior to dataset balancing procedure, the average accuracy across all patients fell within the range of 0.88 to 0.97. These overoptimistic performances include precision, recall, and F1 values around the 0.75 mark, with AUCs values ranging between 0.83 and 0.95. Following data augmentation to reduce the possible bias induced by the data classes imbalance, there was an observable improvement in all performance metrics when assessing the average values. This enhancement suggests an increased level of seizure detection capability.

4.1 Prior to data augmentation

If we take the example of patient number 3132, the k-NN model achieved high accuracy, 0.92 and 0.94 with time-domain features and time-frequency domain features respectively. Precision is around 0.80, but the recall is lower for both sets of features suggesting that the model is affected by the imbalance of the dataset.

The weighted k-NN model's performance results indicate that the model performs very well. In fact, the outcomes showed that different types of features allowed for an effective

identification of epileptic epochs in patient 3132. The values of accuracy of 1 and 0.93 and precision, recall, and F1 scores consistently equal to 1 for time-domain features and 0.77 for frequency and time-frequency domain features suggest the effectiveness of the model in accurately identifying seizure events from audio recordings. Nevertheless, being the time-domain features dataset quite limited especially within the test set, the results obtained with k-NN, decision tree and neural networks are over-optimistic for this specific patient.

On the other hand, the linear SVM model showed limitations, specifically in precision, recall and F1 score when using a limited number of samples as input. This indicates that the model failed to correctly classify any windows as positive (true positives), predicting only negative instances therefore still being able to achieve high accuracy. Instead, when using time-frequency domain features the number of samples is increased, and the linear SVM model was able to correctly classify some positive instances.

This could be attributed to the fact that when using time-domain features the number of observations of the minority class ("seizure") present in the test set is too low for the linear SVM model to distinguish between positive and negative instances. In contrast, the larger number of samples obtained from time-frequency domain features might contain more discriminative information relevant to the studied classes. Imbalanced data can therefore negatively impact the performance of certain models, leading to difficulties in correctly identifying positive instances. Linear SVM might be more sensitive to certain types of features or patterns present in the time-frequency domain feature set, enabling it to achieve better performance compared to the time-domain features. Hence, it might not be well-suited for the specific characteristics of a limited number of test samples (here, those from time-domain features). Moreover, in this study we implemented a linear SVM with a linear kernel, and it's possible that the relationship between the features and the target variable (seizure detection) is not linear. More complex models (polynomial or Gaussian kernels) or feature representations might better capture non-linear relationships. The observed difference in performance between using time-domain features against frequency and time-frequency domain features with the linear SVM model is most likely caused by a difference in the number of samples. Time-domain features consist of a single scalar per audio segment, whereas time-frequency domain features include a number of samples per segment, resulting in a significantly larger feature set for training and testing models.

Similarly to the weighted k-NN model, for patient 3132 the decision tree model exhibited competitive performance, though with some variability across feature sets (1 and 0.74).

Neural network models demonstrated good results with consistent performance across different feature sets for patient 3132. Closely to the k-NN model, the recall obtained with time-frequency domain features resulted in a lower value, again suggesting the negative effect of imbalanced data. When averaging among all patients, the results of accuracy obtained using time-domain features remained quite high for all models in a range between 0.88 and 0.97, whereas the values of precision, recall and F1 score dropped to around 0.80.
Results from the weighted k-NN model suggested good performances with the highest accuracy value for both sets of features. When implementing the method using time-frequency domain features we obtained the highest precision, recall and F1 score compared to the other models.

On average, the decision tree model has returned the best results of precision, recall and F1 score when run with time-domain features, whereas the recall went from 0.74 to 0.58 when using time-frequency domain features. This decrease in recall could be attributed to several factors. Decision tree models make splits in the feature space based on the feature values to create decision boundaries. It's possible that the time-frequency domain features are not as discriminative or informative as the time-domain features for distinguishing between seizure and non-seizure instances. The decision tree might struggle to find meaningful splits in the feature space when using time-frequency domain features, leading to lower recall. Decision trees can easily become complex and overfit the training data, especially if the feature space is high-dimensional or noisy. It's possible that when using time-frequency domain features, the decision tree model overfits the training data, resulting in poorer performance generalization on unseen data and a decrease in recall.

Neural networks models, both with one hidden layer and two hidden layers, obtained the highest values of recall, specifically 0.82, when using time-domain features compared to other models. Instead, with frequency and time-frequency domain features the precision increased from 0.68 (one hidden layer) and 0.70 (two hidden layers) to 0.78, whereas the recall decreased to 0.68. A decrease in recall accompanied by an increase in precision suggests that the model is becoming more conservative in its predictions, correctly identifying fewer positive instances but making fewer false positive errors. This means that it might miss some seizure events (lower recall), which could be problematic if the goal is to detect all seizures accurately to ensure timely intervention or treatment. However, when it does predict a seizure, it's more likely to be correct (higher precision), which can be beneficial for reducing unnecessary interventions or false alarms.

In conclusion, time-domain features, especially when combined with the weighted k-NN and decision tree models, seem to be effective for seizure detection. Neural networks models also

show promise with both sets of features. However, the uniformity of performances equal to 1 across all metrics when utilizing time-domain features for patient 3132, clearly suggests a potential bias induced by the dataset's imbalance and the limited representation associated with these features. There is room for improvement, particularly in models using time-frequency domain features, as they exhibit lower performance metrics. However, the observations associated to these features are more consistent and offer a richer description of the dataset's characteristics.

The evaluation of the Area Under the Curve (AUC) further supported the efficacy of the models in discriminating between seizure and non-seizure instances. While some models achieved perfect AUC values with time-domain features, indicating perfect discriminatory power, others showed slightly lower AUC values with time-frequency domain features. These values suggest that the models can still effectively distinguish between seizure and non-seizure instances based on time-frequency domain features. Therefore, the choice of a feature set can impact model performance in seizure detection tasks.

In conclusion, the AUC results complement the previous performance metrics and provide valuable insights into the discriminatory power of the models, further supporting their potential utility in epilepsy seizure detection tasks. These findings underscore the importance of feature selection and choice of the statistical model in epileptic seizures detection, as well as the potential of neural network architectures for this task since they showed the highest values of AUC for both sets of features.

The range of minimum and maximum values observed across all patients provides insight into the variability of model performance within the dataset. For instance, when considering time-domain features, the weighted k-NN model exhibits a wide range of accuracy, precision, recall, F1 score, and AUC values, suggesting significant fluctuations in its performance across different patients. On the other hand, models utilizing time-frequency domain features generally display narrower ranges, indicating a more consistent performance across the patient population. This suggests that while time-domain features may yield more varied results, time-frequency domain features offer a more stable and reliable performance across different patients.

4.2 Following data augmentation to balance the classes

After data augmentation, we obtained a balanced dataset, with both classes of features evenly represented. Training and testing all machine learning models on a larger and more even dataset contributed to the achievement of better results across all patients and performance metrics.

To illustrate this issue, we considered the results obtained from the recordings of patient 3132. Although we observed lower results when employing methods with time-domain features, these results are more representative and help mitigate the previously mentioned imbalanced-based bias. At the same time, the recall or sensitivity remained equal to 1, suggesting that the models were able to correctly identify all instances of the positive class, indicating a high level of effectiveness in detecting seizures. For this patient, performance metrics increased significantly when utilizing time-frequency domain features, with precision, recall and F1 score values between 0.84 and 0.96. Also, the AUC values were slightly higher than prior to data augmentation.

When averaging across all patients, the values of all performance metrics still resulted higher than prior to data augmentation, especially when implementing time-frequency domain features we observed the greatest increment. This demonstrates how effective data augmentation is in improving the dataset's quality. In fact, the benefits of data augmentation on model performance in seizure detection are highlighted by the improvements seen across precision, recall and F1 score.

It is important to point out that, similarly to the pre-data augmentation scenario, the models demonstrating the highest potential were the weighted k-NN, decision tree, and neural networks. This emphasizes their consistent efficacy across different dataset configurations, highlighting their robustness. Meanwhile, even if linear SVM benefits from data augmentation, it continues to exhibit comparatively lower performance metrics. Although other algorithms such as under-sampling the majority class could be used to balance the dataset, we notice that the reduced size of our dataset strongly limits its use.

The minimum and maximum values observed in the performance metrics across all patients, particularly after data augmentation with SMOTE, reveal notable changes compared to those in the original imbalanced dataset.

For the models employing time-domain features, the range of values appears to be narrower in the augmented dataset, indicating a reduction in variability. For instance, the range for accuracy, precision, recall, F1 score, and AUC in the k-NN model using time-domain features is [0.81, 0.90] for accuracy, [0.77, 1] for precision, [0.91, 1] for recall, [0.84, 0.96] for F1 score, and [0.88, 0.96] for AUC. Comparatively, in the original dataset, the range for the same model was slightly wider, suggesting a more varied performance.

Similarly, when considering models utilizing time-frequency domain features, we observe changes in the range of performance metrics after SMOTE augmentation. The ranges for

accuracy, precision, recall, F1 score, and AUC in various models may differ from their counterparts in the original dataset, indicating potential improvements or alterations in model behaviour due to data augmentation.

Overall, the narrower ranges observed in the performance metrics after SMOTE augmentation suggest a potential improvement in model consistency and reliability.

4.3 Limitations

Detecting seizures from audio recordings presents several limitations, especially with a dataset of limited duration, no presence of total silence and few occurrences of seizures.

The amount of data available for training and testing the seizure detection algorithm is restricted, with only one seizure per patient in a dataset spanning nearly seven hours. The algorithm's capacity to learn and generalize may be compromised by the shortage of seizure epochs. Furthermore, each patient may produce unique acoustic sounds, so a reduced dataset may not adequately capture this variability, limiting any algorithm's ability to generalize across diverse patient populations.

Moreover, the imbalance between seizure and non-seizure instances in the dataset lead to biased model performances. Since seizures are rare events compared to non-seizure periods, the algorithm may not accurately distinguish between the two classes. Seizures can manifest in various forms and intensities, making it challenging to develop a detection algorithm able to recognize all types of seizures. Hence, a study limited to focal and focal to bilateral seizures may not capture the full spectrum of seizure types and characteristics, reducing the algorithm's robustness in real-world scenarios.

Another aspect that is important to keep in mind is that the findings of this study may be influenced by the presence of other sounds within the audio recordings. For example, those categorized as “TV”, “patient’s speech”, and “external voices”, which appear within the “non-seizure” label and can be occasionally detected by the system, may produce false positives. This phenomenon typically occurs during daytime hours, suggesting that the system could potentially perform better during nighttime hours when environmental noise is generally reduced. In addition, artefacts are more likely to occur after an epileptic seizure, which makes it more difficult to identify the episode itself.

Due to the limited number of seizures in the dataset, validating the performance of the seizure detection algorithm and assessing its generalization to new data becomes challenging. While the models might work well with the current dataset, they might not be able to generalize to new or different patient data.

Another limitation is that the process of manually labelling sound events present in the audio recordings can be inaccurate and subjective.

Lastly, the use of windowing with a 50% overlap between windows, potentially resulting in a scenario where 50% of the information has already been exposed to the algorithm during the training phase, could affect the model's ability to generalize effectively.

CHAPTER 5: CONCLUSIONS

This study contributes valuable insights into the field of epilepsy, highlighting the importance of features, statistical models and evaluation metrics. This research contributes to the state of the art in epilepsy seizure detection by utilizing machine learning techniques on audio recordings, which is an area of growing interest and importance in medical research.

While many existing studies focus on seizure detection using electroencephalogram (EEG) or other physiological signals, this study specifically explores the feasibility of detecting seizures from audio recordings. By utilizing audio recording of patients' environment, the study opens up new possibilities for non-invasive and accessible seizure detection methods that may complement existing approaches. The research evaluated two different feature sets, time-domain features and time-frequency domain features, indicating a comprehensive approach to feature engineering.

Our study compared the performance of several machine learning models, including k-NN, linear SVM, decision tree, and neural networks, across both feature sets and both before and after the data augmentation. This comparative analysis helped to identify which models are more effective for seizure detection from audio recordings. The assessment of different performance metrics added depth to the evaluation and provided a more comprehensive understanding of the models' capabilities in seizure detection.

By demonstrating the efficacy of machine learning models in detecting seizures from audio recordings, our study offers insights that could, in the long run, have practical implications for home-monitoring settings. If validated with larger and more diverse datasets, the findings of this study could potentially contribute to the development of real-world seizure detection systems that are non-invasive, cost-effective and accessible to a wider population of patients.

5.1 Future Developments

To overcome the limitations of this study, future developments are required.

Firstly, the dataset will have to be expanded with larger and more diverse samples. This includes additional seizure epochs for each patient, the incorporation of other types of seizures and of new patients. Doing so, it will be possible to create a dataset that is less restricted and more balanced. This could improve the robustness and generalization of seizure detection algorithms.

Investigating additional audio features beyond those currently used in the study could provide a more comprehensive representation of the audio signals. Moreover, feature selection techniques could be used to help identify the most informative and relevant features for seizure

detection, reducing dimensionality and computational complexity while improving model performance. Another future development could be the use of multimodal approaches, for example combining audio data with video data could provide complementary information for more reliable seizure detection. This approach may enhance the sensitivity and specificity of detection models.

Based on the results obtained, further research could also explore more complex neural network architectures (e.g. Long Short Term Memory Networks) to enhance the accuracy and robustness of seizure detection systems based on audio recordings.

BIBLIOGRAPHY

[1] R. S. Fisher *et al.*, "Epileptic seizures and epilepsy: Definitions proposed by the International League Against Epilepsy (ILAE) and the International Bureau for Epilepsy (IBE)," *Epilepsia*, vol. 46, no. 4. 2005. doi: 10.1111/j.0013-9580.2005.66104.x.

[2] "https://www.who.int/news-room/fact-sheets/detail/epilepsy."

[3] A. Van De Vel *et al.*, "Non-EEG seizure-detection systems and potential SUDEP prevention: State of the art," *Seizure*, vol. 22, no. 5. 2013. doi: 10.1016/j.seizure.2013.02.012.

[4] A. Hamlin, E. Kobylarz, J. H. Lever, S. Taylor, and L. Ray, "Assessing the feasibility of detecting epileptic seizures using non-cerebral sensor data," *Comput Biol Med*, vol. 130, 2021, doi: 10.1016/j.compbiomed.2021.104232.

[5] M. C. Smith and J. M. Buelow, "Epilepsy - Introduction," *DM DISEASE-A-MONTH*, vol. 42, no. 11, 1996.

[6] B. F. Shneker, N. B. Fountain, and J. M. Orlowski, "Epilepsy," *Disease-a-Month*, vol. 49, no. 7, pp. 426–478, Jul. 2003, doi: 10.1016/S0011-5029(03)00065-8.

[7] R. S. Fisher *et al.*, "Instruction manual for the ILAE 2017 operational classification of seizure types," *Epilepsia*, vol. 58, no. 4, 2017, doi: 10.1111/epi.13671.

[8] A. T. Berg *et al.*, "Revised terminology and concepts for organization of seizures and epilepsies: Report of the ILAE Commission on Classification and Terminology, 2005-2009," *Epilepsia*, vol. 51, no. 4, 2010, doi: 10.1111/j.1528-1167.2010.02522.x.

[9] Anil Kumar and Sandeep Sharma, *Focal Impaired Awareness Seizure*. 2023.

[10] T. Aung, J. R. Tenney, and A. I. Bagić, "Contributions of Magnetoencephalography to Understanding Mechanisms of Generalized Epilepsies: Blurring the Boundary Between Focal and Generalized Epilepsies?," *Frontiers in Neurology*, vol. 13. 2022. doi: 10.3389/fneur.2022.831546.

[11] P. Ioannou *et al.*, "The burden of epilepsy and unmet need in people with focal seizures," *Brain and Behavior*, vol. 12, no. 9. 2022. doi: 10.1002/brb3.2589.

[12] Z. V. Okudan and Ç. Özkara, "Reflex epilepsy: Triggers and management strategies," *Neuropsychiatric Disease and Treatment*, vol. 14. 2018. doi: 10.2147/NDT.S107669.

[13] M. F. Shaikh, "A Review on Natural Therapy for Seizure Disorders," *Pharm Pharmacol Int J*, vol. 3, no. 2, 2015, doi: 10.15406/ppij.2015.03.00051.

[14] J. A. French, "Refractory epilepsy: Clinical overview," in *Epilepsia*, 2007. doi: 10.1111/j.1528-1167.2007.00992.x.

[15] J. van Andel, R. D. Thijs, A. de Weerd, J. Arends, and F. Leijten, “Non-EEG based ambulatory seizure detection designed for home use: What is available and how will it influence epilepsy care?,” *Epilepsy and Behavior*, vol. 57. 2016. doi: 10.1016/j.yebeh.2016.01.003.

[16] C. E. Elger and C. Hoppe, “Diagnostic challenges in epilepsy: seizure under-reporting and seizure detection,” *The Lancet Neurology*, vol. 17, no. 3. 2018. doi: 10.1016/S1474-4422(18)30038-3.

[17] D. Tax and R. Duin, “Outliers and data descriptions,” *Proceedings of the 7th Annual Conference of the Advanced School for Computing and Imaging*, 2001.

[18] J. Shum and D. Friedman, “Commercially available seizure detection devices: A systematic review,” *Journal of the Neurological Sciences*, vol. 428. 2021. doi: 10.1016/j.jns.2021.117611.

[19] Taeho Kim *et al.*, “Epileptic Seizure Detection and Experimental Treatment: A Review,” *Front Neurol*, vol. 11, no. 701, Jul. 2020.

[20] X. H. Kok, S. A. Imtiaz, and E. Rodriguez-Villegas, “Assessing the Feasibility of Acoustic Based Seizure Detection,” *IEEE Trans Biomed Eng*, vol. 69, no. 7, 2022, doi: 10.1109/TBME.2022.3144634.

[21] S. J. M. Smith, “EEG in the diagnosis, classification, and management of patients with epilepsy,” *Neurology in Practice*, vol. 76, no. 2. 2005. doi: 10.1136/jnnp.2005.069245.

[22] R. S. Fisher *et al.*, “Seizure diaries for clinical research and practice: Limitations and future prospects,” *Epilepsy and Behavior*, vol. 24, no. 3. 2012. doi: 10.1016/j.yebeh.2012.04.128.

[23] P. Ojanen *et al.*, “An integrative method to quantitatively detect nocturnal motor seizures,” *Epilepsy Res*, vol. 169, 2021, doi: 10.1016/j.eplepsyres.2020.106486.

[24] M. N. Istiaq Ahsan, C. Kertesz, A. Mesaros, T. Heittola, A. Knight, and T. Virtanen, “Audio-based epileptic seizure detection,” in *European Signal Processing Conference*, 2019. doi: 10.23919/EUSIPCO.2019.8902840.

[25] Jennifer Shum *et al.*, “Sounds of seizures,” *European Journal of Epilepsy*, vol. 78, pp. 86–90, May 2020.

[26] J. van Andel *et al.*, “Multimodal, automated detection of nocturnal motor seizures at home: Is a reliable seizure detector feasible?,” *Epilepsia Open*, vol. 2, no. 4, 2017, doi: 10.1002/epi4.12076.

[27] C. Carlson, V. Arnedo, M. Cahill, and O. Devinsky, “Detecting nocturnal convulsions: Efficacy of the MP5 monitor,” *Seizure*, vol. 18, no. 3, 2009, doi: 10.1016/j.seizure.2008.08.007.

[28] A. Van de Vel *et al.*, "Non-EEG seizure detection systems and potential SUDEP prevention: State of the art: Review and update," *Seizure*, vol. 41. 2016. doi: 10.1016/j.seizure.2016.07.012.

[29] J. B. Arends *et al.*, "Diagnostic accuracy of audio-based seizure detection in patients with severe epilepsy and an intellectual disability," *Epilepsy and Behavior*, vol. 62, 2016, doi: 10.1016/j.yebeh.2016.06.008.

[30] T. Butko and C. Nadeu, "Feature Selection for Multimodal Acoustic Event Detection," *Group (New York)*, no. June, 2008.

[31] Siqing Wu, "Recognition of Human Emotion in Speech Using Modulation Spectral Features and Support Vector Machines."

[32] T. Heittola, E. Çakir, and T. Virtanen, "The machine learning approach for analysis of sound scenes and events," in *Computational Analysis of Sound Scenes and Events*, 2017. doi: 10.1007/978-3-319-63450-0_2.

[33] A. Abbasi, A. R. R. Javed, A. Yasin, Z. Jalil, N. Kryvinska, and U. Tariq, "A Large-Scale Benchmark Dataset for Anomaly Detection and Rare Event Classification for Audio Forensics," *IEEE Access*, vol. 10, 2022, doi: 10.1109/ACCESS.2022.3166602.

[34] S. A. Syed, M. Rashid, S. Hussain, A. Imtiaz, H. Abid, and H. Zahid, "Inter classifier comparison to detect voice pathologies," *Mathematical Biosciences and Engineering*, vol. 18, no. 3, 2021, doi: 10.3934/mbe.2021114.

[35] M. J. Bianco *et al.*, "Machine learning in acoustics: Theory and applications," *J Acoust Soc Am*, vol. 146, no. 5, 2019, doi: 10.1121/1.5133944.

[36] S. M. Anwar, M. Majid, A. Qayyum, M. Awais, M. Alnowami, and M. K. Khan, "Medical Image Analysis using Convolutional Neural Networks: A Review," *Journal of Medical Systems*, vol. 42, no. 11. 2018. doi: 10.1007/s10916-018-1088-1.

[37] "https://www.axis.com/products/axis-t8351-mk-ii-microphone-35-mm."

[38] I. V. McLoughlin, *Speech and Audio Processing: A MATLAB®-based approach*. 2016. doi: 10.1017/CBO9781316084205.

[39] B. Gold, N. Morgan, and D. Ellis, *Speech and Audio Signal Processing: Processing and Perception of Speech and Music: Second Edition*. 2011. doi: 10.1002/9781118142882.

[40] P. Kathirvel, M. S. Manikandan, S. Senthilkumar, and K. P. Soman, "Noise robust zerocrossing rate computation for audio signal classification," in *TISC 2011 - Proceedings of the 3rd International Conference on Trendz in Information Sciences and Computing*, 2011. doi: 10.1109/TISC.2011.6169086.

[41] S. B. Davis and P. Mermelstein, "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences," *IEEE*

Transactions on Acoustics, Speech, and Signal Processing, vol. 28, no. 4. 1980. doi: 10.1109/TASSP.1980.1163420.

[42] L. Cousyn, V. Navarro, and M. Chavez, "Outliers in clinical symptoms as preictal biomarkers," *Epilepsy Res*, vol. 177, 2021, doi: 10.1016/j.eplepsyres.2021.106774.

[43] I. Brown and C. Mues, "An experimental comparison of classification algorithms for imbalanced credit scoring data sets," *Expert Syst Appl*, vol. 39, no. 3, 2012, doi: 10.1016/j.eswa.2011.09.033.

[44] S. S. Khan and M. G. Madden, "One-class classification: Taxonomy of study and review of techniques," *Knowledge Engineering Review*, vol. 29, no. 3. 2014. doi: 10.1017/S026988891300043X.

[45] S. A. Dudani, "The Distance-Weighted k-Nearest-Neighbor Rule," *IEEE Trans Syst Man Cybern*, vol. SMC-6, no. 4, 1976, doi: 10.1109/TSMC.1976.5408784.

[46] J. Gou, L. Du, Y. Zhang, and T. Xiong, "A new distance-weighted k-nearest neighbor classifier," *Journal of Information and Computational Science*, vol. 9, no. 6, 2012.

[47] Z. Zhang, "A gentle introduction to artificial neural networks," *Ann Transl Med*, vol. 4, no. 19, 2016, doi: 10.21037/atm.2016.06.20.

[48] J. F. E. IV, D. Michie, D. J. Spiegelhalter, and C. C. Taylor, "Machine Learning, Neural, and Statistical Classification.," *J Am Stat Assoc*, vol. 91, no. 433, 1996, doi: 10.2307/2291432.

[49] N. V. Chawla, "Data Mining for Imbalanced Datasets: An Overview," in *Data Mining and Knowledge Discovery Handbook*, 2009. doi: 10.1007/978-0-387-09823-4_45.

[50] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Trans Knowl Data Eng*, vol. 21, no. 9, 2009, doi: 10.1109/TKDE.2008.239.

[51] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, 2002, doi: 10.1613/jair.953.

[52] K. S. WOODS, C. C. DOSS, K. W. BOWYER, J. L. SOLKA, C. E. PRIEBE, and W. P. KEGELMEYER, "COMPARATIVE EVALUATION OF PATTERN RECOGNITION TECHNIQUES FOR DETECTION OF MICROCALCIFICATIONS IN MAMMOGRAPHY," *Intern J Pattern Recognit Artif Intell*, vol. 07, no. 06, 1993, doi: 10.1142/s0218001493000698.

[53] M. Khushi *et al.*, "A Comparative Performance Analysis of Data Resampling Methods on Imbalance Medical Data," *IEEE Access*, vol. 9, 2021, doi: 10.1109/ACCESS.2021.3102399.

[54] G. M. Weiss and F. Provost, “Learning when training data are costly: The effect of class distribution on tree induction,” *Journal of Artificial Intelligence Research*, vol. 19, 2003, doi: 10.1613/jair.1199.

[55] N. Japkowicz and S. Stephen, “The class imbalance problem: A systematic study,” *Intelligent Data Analysis*, vol. 6, no. 5, 2002, doi: 10.3233/ida-2002-6504.

[56] G. E. A. P. A. Batista, R. C. Prati, and M. C. Monard, “A study of the behavior of several methods for balancing machine learning training data,” *ACM SIGKDD Explorations Newsletter*, vol. 6, no. 1, 2004, doi: 10.1145/1007730.1007735.

[57] Alexander Yun-chung, “The Effect of Oversampling and Undersampling on Classifying Imbalanced Text Datasets,” *CWL Publishing Enterprises, Inc., Madison*, vol. 2004, no. August, 2004.

[58] A. Amin *et al.*, “Comparing Oversampling Techniques to Handle the Class Imbalance Problem: A Customer Churn Prediction Case Study,” *IEEE Access*, vol. 4, 2016, doi: 10.1109/ACCESS.2016.2619719.

[59] Z. Zheng, Y. Cai, and Y. Li, “Oversampling method for imbalanced classification,” *Computing and Informatics*, vol. 34, no. 5, 2015.