JOHANNES KEPLER
UNIVERSITY LINZ

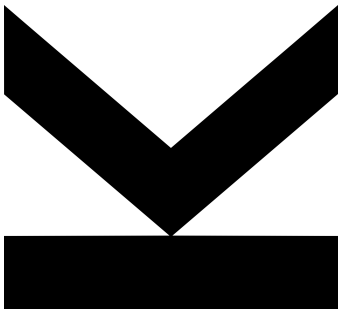
Author
Alexander Wolf, B.Sc.
k11725625

Submission
Institute for Business Informatics - Software Engineering

Thesis Supervisor
DI Dr. **Wolfgang Narzt**

May 2025

Prediction of epileptic seizures using volatile organic compound data



Master Thesis

to obtain the academic degree of

Master of Science

in the Master's Program

Business Informatics

JOHANNES KEPLER UNIVERSITY LINZ
Altenbergerstraße 69
4040 Linz, Austria
www.jku.at
DVR 0093696

Abstract

Epileptic seizures can cause serious injuries when individuals with epilepsy are unable to move to a safe position in time. Predicting seizures in advance is therefore crucial for ensuring patient safety and reducing costs associated with seizure-related injuries. Dogs, known for their highly sensitive noses, have demonstrated the ability to accurately predict epileptic seizures, prompting the question of whether a wearable device could replicate this capability. This study investigates the feasibility of using electronic noses (e-noses) combined with machine learning frameworks for seizure prediction. In the EPILEPSIA study, smell data was collected from individuals with epilepsy using e-noses. Given the numerous hyperparameters involved in seizure prediction, a grid search was conducted to explore these options. The grid search trained 576 time-series CNN classifiers, yielding promising results, with the best model achieving an accuracy of 77%. However, the small dataset limited the robustness of these findings, as feature permutation and noise tests revealed instability and potential overfitting of the model. These results emphasize the need for further research with larger datasets to validate the potential of wearable e-nose devices as reliable tools for seizure prediction.

Contents

List of Figures

List of Tables

1 Introduction

Imagine living with the constant uncertainty of when the next seizure will occur. You could be walking down a busy street, standing at the edge of a train platform, or simply preparing dinner in your kitchen. Without warning, a seizure could take place, leaving you vulnerable to serious injury or worse. For millions of people with epilepsy, this is a daily reality — a life shaped by unpredictability and fear. Beyond the physical dangers, this uncertainty takes a toll on mental health and overall quality of life. It forces individuals to plan their lives around a condition they cannot control. Despite advances in medicine and technology, the ability to predict a seizure in their daily life remains an unmet need. What if there were a way to predict these events before they happened? This question drives a growing body of research aimed at creating innovative solutions to bring safety and peace of mind to those living with epilepsy.

1.1 Problem Statement

Epilepsy is one of the most prevalent brain conditions affecting over 50 million people on earth. Premature death is nearly three times as likely for people with epilepsy. On top of physical harm that the neurological condition can cause, people with epilepsy are likely to be stigmatized and discriminated against. There also exists a significant treatment gap, as low-income countries have more cases of epilepsy yet do not have the means to treat them. While there are remedies available that can make people with epilepsy seizure free, helping people with epilepsy will need a multifaceted approach. [1, 2]

Another approach, apart from trying to make people seizure free, is to know beforehand if a person is about to have a seizure and warn them in time. It has been shown that with data from specialized equipment, such as an electroencephalogram (EEG), machine learning models can be trained to do just that with adequate accuracy [3, 4]. While this is

an incredible achievement, it is crucial to remember that this prediction can only be done in a clinical setting, due to the size of the equipment involved. There is currently no way for people with epilepsy to be warned in time when going about their daily lives.

The ability to predict seizures in real-time would allow individuals to take necessary precautions, avoid dangerous situations, and potentially reduce the risk of injury. Moreover, it would provide a sense of control and peace of mind, significantly improving their overall well-being. [5, 6]

1.2 Learning from Canine Abilities

Dogs have an extraordinary sense of smell, with a very low threshold for olfactory detection [7]. While humans are only able to smell a substance if it can be measured at 40-60 parts per billion (1×10^{-12}), dogs can go as far as smelling compounds in the range of parts per trillion (1×10^{-15}) [8]. This extraordinary olfactory ability allows them to detect minute changes in the chemical composition of sweat, which can also occur before a seizure. This ability in combination with operant conditioning allows for dogs to be trained to identify all kinds of diseases. Not only can they be trained on sweat samples, but also on saliva, urine or breath samples [9]. While in other industries trained dogs have always played an important role, for example sniffer dogs for police or the army, this level of involvement of dogs is not matched in the medical field. Doctors and medical experts mostly prefer advanced technical methods instead, although dogs often deliver accuracy rates on par with technical approaches. It is most likely that the adoption rate is low due to the novelty and less standardized nature of olfactory detection by dogs. [9]

The exact mechanism by which dogs detect these changes is not yet fully understood, but it is believed that they can sense specific smells that are released by the body in response to certain diseases. In the case of epilepsy, electrical brain activity is not the only biomarker that is changing before a seizure occurs. There have been numerous studies observing dogs correctly distinguishing sweat samples before and after a seizure of a person with epilepsy [10, 11, 12, 13]. It can therefore be concluded, that there are volatile organic compounds being released before a seizure that are definite indicators that a seizure is about to happen. By collecting smell data of people with epilepsy, one should be able to reproduce the same result, given the correct volatile organic compounds are being

measured. The difference of course being, that by devising a correct machine learning algorithm and packing it into a wearable device, people with epilepsy could be warned of upcoming seizures.

While we talk about the different smells that dogs pick up, these smells are specific volatile organic compounds (VOCs), which are chemical substances that easily evaporate at room temperature and can be identified in the air. These compounds are released by the human body and may change in response to physiological conditions, such as an impending seizure.

The potential for using volatile organic compound data as a biomarker for seizure prediction is significant. If these compounds can be reliably identified and measured, it could lead to the development of non-invasive wearable devices that monitor VOC levels in real-time. Such devices could provide early warnings of impending seizures, allowing individuals to take preventive measures and avoid potentially dangerous situations.

1.3 Scope and Delimitations

The goal of trying to predict epileptic seizures is an enormously ambitious one, which is why the scope of this master thesis has to be limited to ensure that the objectives are achievable within the given timeframe and resources. This thesis will focus specifically on the analysis of volatile organic compound data collected through a smell sensor. Other aspects, such as the integration of additional sensor data or the development of a comprehensive seizure prediction system, will not be covered in this work. The primary aim is to explore the feasibility of using volatile organic compounds (VOCs) as biomarkers for seizure prediction and to develop a preliminary machine learning model based on this data. Furthermore, a closer look at the impact of hyperparameters will be taken.

The machine learning framework SKTime and specifically the convolutional neural network (CNN) classifier [14], which has been specifically designed for time-series data, will be used to build experimental machine learning models with the collected data.

The variation in the parameter grid that will be examined will be limited due to computing power and time constraints. For each fixed variable in the grid-search, reasoning for not exploring the option further will be provided.

1.4 Research Questions

The main focus of this master thesis will be on analysing the data gathered through a smell sensor and afterward trying to predict epileptic seizures with this data. Even though the likelihood of predicting epileptic seizures increases with the number of different features being incorporated in the training of a machine learning model, it is not the goal of this thesis to predict seizures based on the whole plethora of vital parameters available. The task of combining all features into one prediction model would be out of scope for this master thesis.

Therefore, we come to the conclusion that this master thesis will try to answer the following research question:

1. How effectively can epileptic seizures be predicted using volatile organic compound data combined with a convolutional neural network and grid-search optimization?

1.5 Methodology

A general understanding of epilepsy as a condition as well as volatile organic compounds is necessary for all later chapters, which is why chapter 2 gives an overview over both topics. The related work chapter showcases what has already been tried in epileptic seizure prediction research and where the research gap currently is. In chapter 4 we start of by explaining where and how we collected the data used in the later analysis. Afterward, all possible hyperparameters that could meaningfully affect model performance will be discussed. Since there is no comparative literature that can suggest optimal values for our context-specific hyperparameters, a grid-search of possibly impactful settings for the CNN model will be done. All training is done using a set seed, as to guarantee reproducibility. In the grid-search we trained 576 differently configured models, gauging what performance is possible using our collected data. Label shuffling tests were conducted to make sure well

performing models are not doing so by chance. Lastly, feature permutation, perturbation and ablation tests will assess the trained model regarding its robustness.

1.6 Terminology

In this thesis I will refrain from using the term "epileptic person" or using the abbreviation "PWE" (person with epilepsy), since both are considered stigmatizing and have a negative societal connotation [15, 16, 17].

2 Foundations

2.1 Epilepsy

To understand why the goal of predicting epileptic seizures ahead of time is meaningful, it is imperative to understand in which way epilepsy manifests itself and how people living with epilepsy are being affected by it. For this reason this chapter will give a broad overview about the theoretical and practical aspects of the condition. To make for a more coherent introduction to the condition we will begin by examining the prevalence of epilepsy and continue by delving into the possible causes of how the condition is developed. Afterwards we will take a close look at the different types of epilepsy and how they are currently being classified. Following the categorization we will discuss the different phases of an epileptic seizure and which treatments are commonly being given.

It goes without saying that we will only be scratching the surface of the current literature and not all details will be covered. This chapter simply provides an introductory overview as many theoretical concepts will come into play when examining the gathered data from our study.

2.1.1 Epidemiology

Epidemiology studies the frequency of occurrence and distribution of a condition in a portion of the population. In the case of epilepsy there currently approximately 50 million people affected worldwide with an incidence of 5.4 to 8.1 per 1,000 people per year [18]. The share of the population with epilepsy will likely increase over time. The reason for this is twofold: epilepsy often occurs as a subsequent condition after surviving a serious accident or stroke, secondly, due to the increased median life expectancy [19], humans

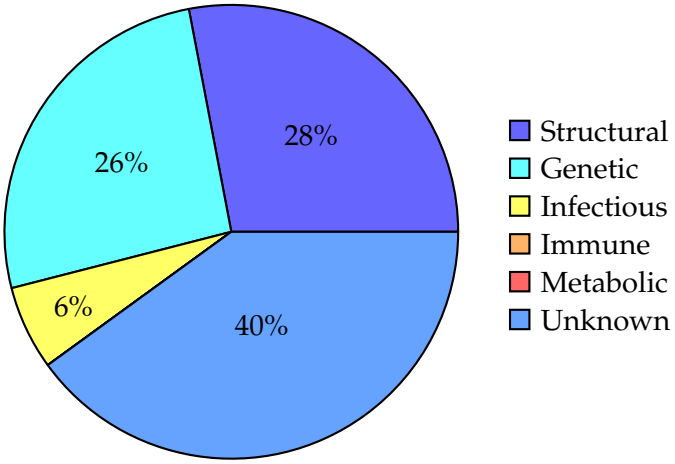
are reaching ages in which the chance of developing epilepsy increases significantly. The chances for developing epilepsy are highest at the ages 5 to 9 and 80 and above. While age does play an important role in the prevalence of epilepsy, gender does not. [20]

Developing countries or countries with higher parasitic infection rates, subpar health care systems or less access to anti-seizure drugs report an up to three times higher prevalence of epilepsy than developed countries [20, 21]. Estimates suggest that 70% of current active epilepsy (AE) cases could live seizure free if treated correctly. [22]

2.1.2 Etiology

Etiology studies the origin of a condition. In the case of epilepsy for a major portion of all people with epilepsy a definite causality cannot be determined. Epilepsy has countless risk factors which have to be accounted for when determining the origin of the condition. The most impactful factors being injury severity, low grade tumours, Alzheimer's disease and strokes. Alcohol consumption played less of a role in epilepsy development. Just as the prevalence differs from country to country, so do the etiological reasons for the development of the condition. Risk factors such as perinatal trauma and infections are more likely to occur in low income countries. [23]

Figure 2.1: Etiology of adults with epilepsy between the ages 18-84. [24] There were no cases with metabolic related etiology and only 2 cases of the total 653 had immune related etiological causation.



The distribution seen in section 2.1.2 shows the etiological origins for epilepsy in a sample group of adults. Structural etiology refers to abnormalities visible on neuroimaging, with which clinicians can visually determine the cause of seizures to be the visible abnormality. More clearly, genetic reasons determine the origin of the condition to be a known or presumed mutation, not all genes involved in this process are known yet. Infectious reasons stem from seizures being a symptom for many diseases, such as meningitis or encephalitis. It is important to note that according to the International League Against Epilepsy (ILAE) classification a patient can fall into more than one etiological category. [25]

2.1.3 Categorization

Categorization of cases of epilepsy, is a vital part of the research concerning epilepsy as it functions as the foundation for further research and treatment. The ILAE classification in and of itself is an important clinical guideline on how to categorize epilepsy cases. Subsequent research and treatment always depends on the current diagnosis, which, if imprecise, leads to worse outcomes for patient and research alike. Therefore, the specification files each case according to the following categories. This categorization is independent of the etiological diagnosis. The following categorization based on the ILAE classification of the epilepsies is a first step in categorizing epilepsy cases.

- Seizure type
  - Focal onset
  - Generalized onset
  - Unknown onset
- Epilepsy type
  - Focal
  - Generalized
  - Combined Generalized & Focal
  - Unknown

The diagnosis according to the classification is to be seen as building upon each other. The precondition being that the clinician already ruled out all diagnosis other than epilepsy. It is possible to only be able to classify according to the first level, seizure type, particularly if a patient only experienced a single seizure.

Statistically speaking, some types of seizures and epilepsies are more prone to developing learning disabilities or can be at a higher risk of other comorbidities such as Sudden Unexpected Death in Epilepsy (SUDEP). This again underlines the importance of a correct classification.

The ILAE 2017 specification also refers to a third level, classification by epilepsy syndrome, however a formal list of possible syndromes is not provided. The specification paper refers to their website [26], which contains an extensive list of syndromes. These syndromes are commonly known occurrences of epilepsy with a certain combination of seizure and epilepsy types. [25]

As we begin speaking about focal and generalized seizures it is important to understand the difference between the two. One way to differentiate between focal and generalized seizures is by looking at the electroclinical markers, by means of an EEG. While generalized seizures have their origin at one specific point in the brain and activate bilaterally distributed networks, focal seizures are limited to one hemisphere in the brain. [27]

While the initial classification paper is a useful tool for clinicians to make a first categorization, the operational classification guideline goes into detail about how one can determine a certain seizure type [28]. Epileptic seizures can present themselves with wildly different symptoms. Distinctions can for example be made when looking at the awareness of the patient during the seizure. Patients with generalized seizures typically lose their awareness completely during the seizure or have little memory of the event. In the case of focal seizures, patients retain their awareness at the onset of the seizure, even though they sometimes lose awareness later on. The distinction between characteristics of seizures can be beneficial for machine learning experiments with only certain seizure types.

Figure 2.2 shows the full operational classification of seizure types. While there are many specific categories a seizure can fall into, the most important distinction is based on the motor component. There are many ways a motor onset can present itself, whether through jerking motions (clonic), muscles becoming limp (atonic) or tense and rigid muscles (tonic).

Non-motor onset on the other hand may show itself through behaviour arrest, such as a staring spells. One interesting type of motor onset that should be highlighted is in the form of automatism. In this case the pre-ictal movement is continued throughout the seizure.

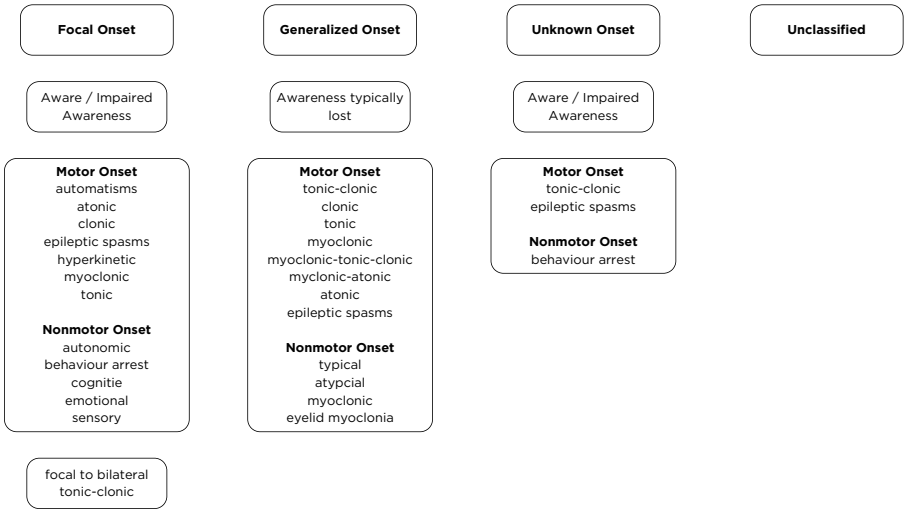


Figure 2.2: Operational classification of seizures types according to the ILAE.

2.1.4 Phases of an epileptic seizure

What is especially important, also for our task, is the distinction that is to be made between different phases of a seizure. As each phase is characterized by different symptoms and physiological changes. Understanding these phases is crucial for both diagnosis, treatment and possible prediction. While there multiple ways to differentiate between seizure phases, the one most useful for our use case will be elaborated on.

2.1.4.1 Pre-ictal phase

This phase is also sometimes called prodromal or aura phase. While there are varying opinions, research defines the pre-ictal phase as the time window few minutes before the seizure begins. This will be especially important later on, as the variation of the picked

time window will impact machine learning results greatly. Le Van Quyen et al. identifies the time window 5 minutes before the seizure as the pre-ictal phase [29]. The pre-ictal phase is the ideal phase where a prediction of future seizures can be made. The patient may experience subtle changes in mood, behaviour, or sensations during this phase, but is not experiencing any symptoms of an actual seizure. Not all patients can tell when they are in a pre-ictal state, since this would mean that all patients could themselves predict when a seizure will occur.

2.1.4.2 Ictal phase

The ictal phase is the seizure itself, defined as the peak of abnormal electrical activity in the brain. Behavioural changes may present themselves during the ictal state, meaning there is not only one way to detect a seizure [30]. This phase can vary greatly in duration and intensity, depending on the type of seizure. The symptoms and types of seizures that can take form have been explained in section 2.1.3.

It can sometimes be difficult to differentiate between inter-ictal abnormal activities in the brain and actual ictal states. [31]

2.1.4.3 Post-ictal Phase

The post-ictal phase follows the ictal phase and can last from minutes to hours to days. During this phase, the brain is recovering from the abnormal electrical activity. Individuals may experience confusion, drowsiness, headache, and memory loss. The severity and duration of the postictal phase can vary depending on the type and duration of the seizure. [32]

2.1.4.4 Inter-ictal Phase

The inter-ictal phase is defined as any time between ictal or post-ictal states. Since it is difficult to pinpoint the exact end of a post-ictal state, this is also the cases for inter-ictal phases since they are adjacent. [31]

2.1.5 Treatment options

There are drug-free ways patients can become seizure-free, for example through hormonal therapies, diet, surgery, neurostimulation, and behavioural modification techniques. While these have been proven to work to some degree, the most effective way to treat epilepsy is by using Anti-Epileptic Drugs (AED).

While effective, patients can experience serious side effects that need to be considered when choosing an AED. Treatment options need to be evaluated over time, since it is rare for patients to successfully receive the same treatment over the years. [33]

Even with treatment, some patients will not be able to live seizure-free, making a prediction system even more valuable.

2.2 Volatile Organic Compounds

2.2.1 Definition

Volatile Organic Compounds (VOCs) are a large group of carbon-based chemicals that easily evaporate at room temperature. They are called volatile because they have a high vapour pressure at ordinary room temperature, which means they can easily become vapours or gases. All living beings, including humans, emit volatile organic compounds. These emissions are usually coming from breath, sweat from the skin, urine or blood itself. The VOCs that are emitted by a human can be used to make certain assumptions about them regarding their health [34]. There is an important distinction to be made in the origin of a volatile organic compound. VOCs can also originate from anthropogenic sources, such as the processing or burning of fossil fuels or the evaporation of solvents used in an industrial complex. Even though the VOCs coming from anthropogenic sources often times cause adversary health issues, they are not of interest for this theoretical elaboration. [35, 36]

2.2.2 Measurement Techniques

The standard way to measure VOCs has been gas chromatography mass spectrometry (GC/MS), while accurate, this method has been deemed as slow and expensive, which is why multiple new ways for separating and detecting VOCs have emerged in the last few years [37]. Other methods like proton-transfer reaction mass offers a faster response time and are equally sensitive, which is why they are used in the detection of trace levels of VOCs in the atmosphere [38]. Selected ion flow tube mass spectrometry (SIFT-MS) has also been proven to work in situations where real-time detection of VOCs is required and a wide range of VOCs needs to be covered [39].

In our context, the most important way of measuring VOCs in the air is the electronic nose. Different to the methods like GC/MS or SIFT-MS, where the measured VOCs can be specified, e-noses work with a sensor array of non-selective gas sensors. These non-selective gas sensors only detect the presence and concentration of VOCs but do not specify which VOCs are currently being measured. As a single sensor is completely non-selective, no valuable information could be learnt from it. When bundled in an array of sensors, which all respond slightly differently to the VOCs around them, a fingerprint of different VOCs can be captured. Through pattern recognition algorithms the presence of individual VOCs can then be captured. [40]

2.2.3 Factors influencing Volatile Organic Compound Profiles of Humans

Humans emit a wide range of different VOCs in different amounts. They emit so many that it is in fact possible to uniquely identify each human being based on their VOC profile [41]. Even more important than unique identification, the individuals VOC profile gives insight into their current well-being. This insight is only possible if the recording of the VOC profile goes undisturbed and is not contaminated with unrelated data. Depending on from where the VOC data is being captured this poses more or less of a risk. Wang et al. showed that environment conditions such as temperature or even the clothing of an individual can affect the amount and concentration of the VOCs being picked up by the sensor [42]. Smoking behaviour as well as age, BMI or gender can have an effect on the human VOC profile [43], this could be especially relevant for future prediction tasks using VOC data. In an experiment in a closed classroom, Tang et al. found personal care

products to be a massive emitter of VOCs as well as human skin oil oxidation by ozone present in the air [44].

Inhibiting factors like clothing or masking factors like personal care products need to be addressed when experimenting with VOC-based diagnosis as these factors could disrupt clean data collection and possibly lead to false results.

3 Related work

In chapter 2 the theoretical aspects of both epilepsy and volatile organic compounds have already been discussed. Therefore, in this related work chapter the literature regarding epileptic seizure prediction and prediction of other diseases using VOC data will be highlighted.

3.1 Status quo of epileptic seizure prediction

Gaisberger wrote his master thesis on the status quo of seizure prediction, considering all types of biomarkers and machine learning techniques [45]. This work has shown that often times studies regarding seizure prediction and detection do not classify the type of seizure. Cases are then broadly classified as "epileptic seizures", even though there exists a myriad of different seizure types, which could serve as additional and critical meta information about the study and the results. The reasoning behind this could be, that this field of study is still in its infancy and considering all types of seizures when predicting or classifying data broadens the data set. Of course, from a research perspective clearly defining which types of seizures are easy or harder to predict and what caveats come with each one is still interesting.

In addition to the conclusions about seizure classification, [45] also showed that EEG data is the number one factor when it comes to classification as well as prediction, with over 85% of studies in the meta analysis using EEG data for prediction and detection. There are other biomarkers that are currently being used, but mostly in combination with the EEG data available. Heart-rate variability, acceleration data and electrocardiography (ECG) data, blood volume pulse and electrodermal activity data (EDA) were infrequently used, only in up to 4% of the studies.

Regarding the distribution of machine learning methods used in the detection and prediction of epileptic seizures, no clear favourite can be determined according to [45]. All currently popular techniques, such as CNN, ANN, RNN and LSTM were used in the studies and performed well.

3.2 Volatile organic compound based detection of epilepsy

The idea of using olfactory data to make a distinction between people with and without diseases is not new. This is also the case for epilepsy. Dartel et al. showed that e-noses were able to distinguish breath profiles of people with epilepsy and a control group of people without epilepsy. The prediction model created in the study reached a sensitivity of 76%, specificity of 67% and an accuracy of 71%. The study disclaimed that the results were subpar, due to the use of anti-epileptic drugs in their participants. This suspicion had been confirmed by assessing two additional control groups, with and without anti-epileptic drugs. [46]

3.3 Canine seizure prediction

While e-noses have not been in use regarding the prediction of seizures, dogs have been known to show a reaction before seizures. There have been several studies proving that there is a specific seizure odour that is emitted before the seizure, which the sensitive dog noses are able to pick up. [13] clearly showed that, dogs cannot only smell odours of other diseases such as breast or lung cancer, but smell upcoming epileptic seizures as well. [12] shows that even untrained dogs show a reaction before the onset of a seizure.

While showing that dogs have these capabilities is already a huge step forward, finding out what exact VOC they are smelling when correctly identifying a seizure is even more relevant for our use-case. In a 2016 dissertation project Davis for the first time identified menthone as a possible biomarker in the pre-ictal phase of epileptic seizures [47]. These results have then been confirmed in later studies by Maa et al., where menthone was also identified as one of the primary VOCs found in pre-ictal people with epilepsy [10]. Additionally, canine trials showed that the olfactory prediction of seizures precedes all

electronic biomarkers by "a considerable amount of time", according to Maa et al. Additional VOCs that could potentially be of relevance, such as menthyl acetate or camphor, are also listed. In a follow-up study Maa, Arnold, and Bush further cemented menthone as a primary component of seizure-scented sweat. However, this study revealed that this VOC cannot only be found in people with epilepsy, but can be found in all humans when they experience fear. This was shown, as dogs were not able to tell the difference between seizure sweat samples and fear sweat samples [11]. While there are now multiple studies stating menthone as a VOC, it is important to not rule out any other possible VOCs or patterns in this data analysis.

3.4 Volatile organic compound based prediction of other diseases

At the time of writing, there are no studies trying to predict or detect epileptic seizures using VOC data. Contrarily, there have been multiple attempts to detect various forms of cancer using smell data gathered by e-noses [48, 49, 50, 51]. All of these studies showed promising results with accuracy, sensitivity and specificity upwards of 85% using LDA, logistic regression and kNN.

VOC data has not only been used to predict cancer, but also chondrosarcoma [52], carcinoma [53] and candidemia [54]. All of these studies showed promising results, which underlines the importance of possible similar results for seizure prediction using VOC data.

4 Approach

4.1 The EPILEPSIA project

Before going into the preprocessing and data analysis it is important to elaborate on how the data has been collected. This master thesis is made possible due to the data collected in the EPILEPSIA study, which was conducted at the Neuromed Campus and the Med Campus IV in Linz. The primary goal of this study was to collect as much vital parameter data of persons with epilepsy as possible. By default, persons with epilepsy were regularly scheduled for checkups in the video-EEG units in the two respective hospitals in Linz, where they remained on average for 2-7 days. Upon arrival, they were asked if they wanted to participate in our study, which would mean being connected to an additional set of sensors. These additional sensors and their placement can be seen in Figure 4.1. In Table 4.1 the different vital parameters which are being collected by the respective sensors can be seen. It is important to note that these sensors did not interfere in any way with the regular EEG that the person with epilepsy is required to wear during the checkup. Meaning that neither placement of EEG-electrodes nor recorded data was disturbed by our study.

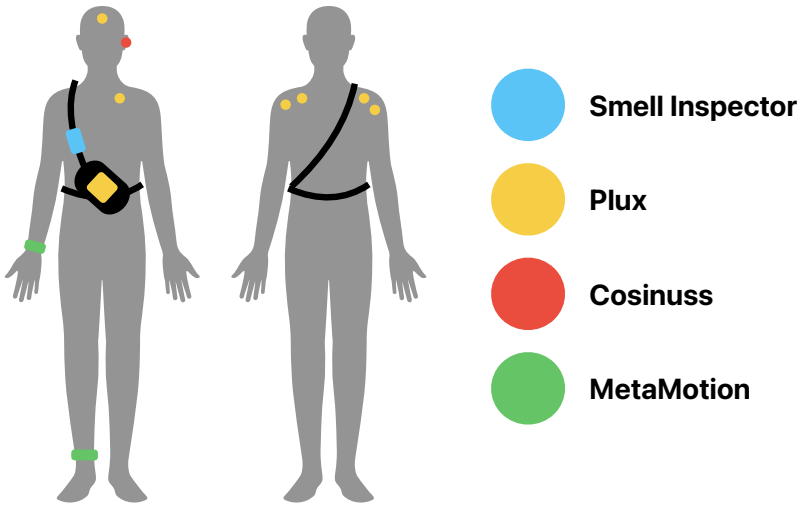


Figure 4.1: Front view (left) and back view (right) of the EPILEPSIA study setup.

Sensor	Vital parameters
Cosinuss	Photoplethysmogram, blood oxygen levels, heart-rate, temperature, respiration rate, perfusion
MetaMotion	Accelerometer, Magnetometer, Gyroscope
Plux	Electromyography, electrodermal activity, near-infrared spectroscopy
Smell Inspector	64-channel sensor array reacting to different VOCs

Table 4.1: Used sensors and the vital parameters they record

The Smell Inspector, was deliberately placed near the armpit where the most VOC data could possibly be recorded. This was decided after consulting with doctors at JKU and technical staff at Smart Nanotubes Technologies, the company behind the Smell Inspector.

In addition to the sensors themselves, there were many hardware and software components necessary during our study. These components were mostly used by medical staff whenever the sensors of a patient went low on battery and needed to be switched. Here is a short overview over all components that were used during the study.

- Android tablet app for connecting sensors and sending data to the Influx database
- Web-app dashboard giving an overview over current patients and sensors
- Influx database for storing sensor data
- MongoDB database for storing metadata

When a sensor ran out of battery, all five sensors were switched out collectively, as to guarantee a regular sensor swapping cycle. The sensor with the shortest battery life was the Plux sensor, with 12 hours battery, meaning every 12 hours all sensors needed to be changed. This was done by medicine students that were paid to help conduct the study.

While there were technical and organizational issues from time to time, leading to some data loss, a satisfactory amount of data has been collected for each patient. We expected some technical issues, since the Bluetooth connection between the sensors and the tablet cannot be guaranteed to be completely stable.

4.2 Pipeline architecture

In order to get to a valid result, a lot of detailed processing has to be done. To keep the end goal in mind and not get lost in the details, this short overview should be used as a reference which steps are going to be taken. The overview of the pipeline architecture that has been created in this prospective study can be seen in fig. 4.2. In all later chapters a more detailed view of the pipeline will be displayed.

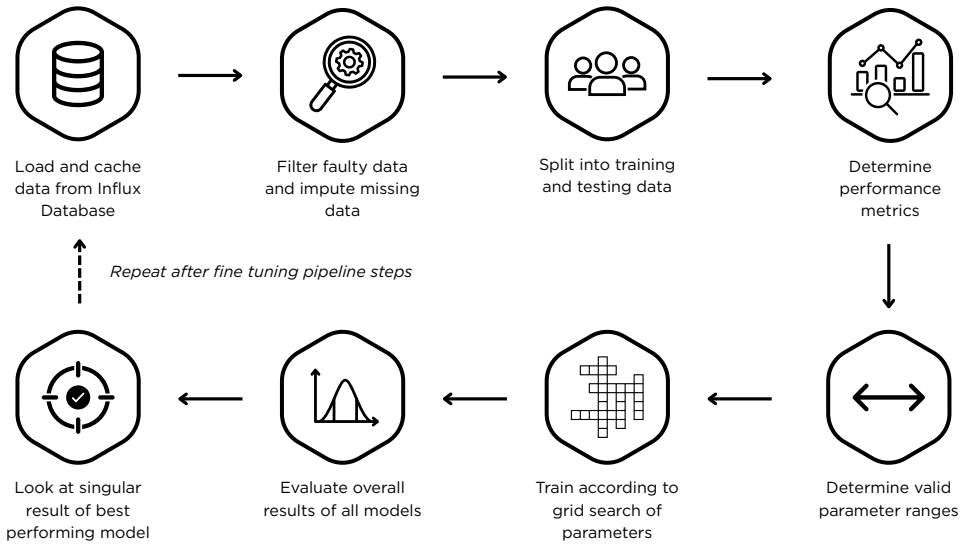


Figure 4.2: Overview of the pipeline architecture.

The pipeline begins by loading and caching data from the Influx Database, which is built for working with time-series data. Caching ensures that repeated data access during model development is efficient and does not require constant querying of the database. Once the data is available, the next crucial step is cleaning it. The issue of faulty or missing data points need to be addressed, which is why the next chapter is of utmost importance.

Afterwards, the data is split into training and testing sets according to a specific strategy that allows for the best results in our use case. Before starting the grid-search, two important variables need to be determined. Firstly, according to which parameters should the trained models be ranked and secondly, what hyperparameter ranges should be included in the grid-search.

Once training is complete, the results of all models are evaluated to identify the most promising configurations. From this group, the single best-performing model is examined in detail to understand its predictions and behaviour. Finally, based on these insights,

the pipeline can be fine-tuned and the process repeated, improving each stage based on lessons learned in the previous iteration.

4.3 Preprocessing

Preprocessing the collected data poses a significant challenge in our case, as there are countless ways to process the data, but limited research to guide the decision-making process. This uncertainty is one of the reasons a grid search was conducted, as it helps identify the options that yield the best results.

Many preprocessing choices, while not typically considered hyperparameters, can have a substantial impact on model training. As a result, these options, such as the length of the prediction window, are included in the grid search. For example, the grid search might compare prediction windows of 5 minutes and 10 minutes to assess how varying this parameter affects performance. All preprocessing choices treated as hyperparameters are detailed in section 4.4.

To give an overview of the following steps beforehand, fig. 4.3 shows what tasks need to be accomplished, before any analysis can be done on the data.

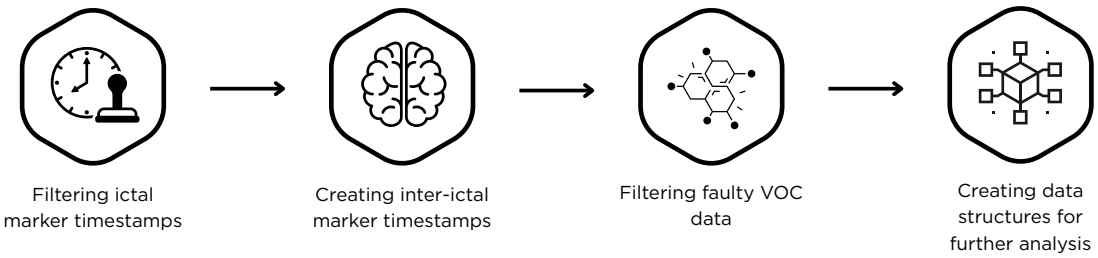


Figure 4.3: Overview of the preprocessing steps within the pipeline.

4.3.1 Marker dataset

Markers, in our case, are the timestamps that were given to us by the medical staff that denote the start and end of a seizure for a given patient. Having these timestamps is the crucial element upon which all further analysis is based upon. Given these timestamps we can calculate the start of a time window that will be feed into a machine learning framework. For our purpose, the most important differentiation should be made between ictal markers and inter-ictal markers. Ictal markers pinpoint the start of a seizure, while inter-ictal markers are used to identify time windows which can be used as baseline. Only these two categories are necessary, because pre- and post-ictal markers can simply be calculated based on the initial ictal marker.

In total, we have gathered data from 81 people with epilepsy of which 30 had seizures while participating in our study. For these 30 people with epilepsy we have 352 markers denoting either the starts or ends of seizures during their time participating in the study.

Ictal marker dataset

Knowing the exact moment a seizure occurred is the most important factor when analysing seizure time series data. For that reason, seizure markers corresponding to the start and end times of seizures of all patients are our basis for further analysis. These seizure markers were manually created by doctors in the NeuroMed and MC IV Campus.

We have the following columns of information about each of the 352 markers in our database.

- EPILEPSIA ID
- Time
- Marker
- Type

The EPILEPSIA ID is a 6 digit combination of upper case letters and numbers to uniquely and identify each patient in the study. This ID is pseudo-anonymous, meaning that

non-hospital staff cannot retrace which EPILEPSIA ID belongs to which person with epilepsy.

The marker column contains the respective start and end markers for the video and EEG signal. In Table 4.2 the respective counts of each marker are displayed.

Marker	**Count**
EEG Seizure Start	96
EEG Seizure End	104
Video Seizure Start	71
Video Seizure End	81

Table 4.2: Counts of the different seizure markers obtained.

Since all of our study participants were observed in Video-EEG units there are two different methods of confirming a seizure. Firstly, often times the seizure causes a motor reaction, causing the person to twitch or stiffen their muscles. The start and end of this reaction can then be seen on the video recording and therefore the timing can be deduced. Secondly, trained professionals and doctors are able to determine a seizure start and end based on an EEG recording. A non-motoric seizure cannot be seen visually and there might only be the EEG signal to correctly identify a seizure start and end.

Due to human error, not every seizure is correctly labelled with a corresponding start and end marker. There are cases where start or end markers are non-existent for a given seizure. The discrepancies are displayed in Table 4.3. For the purpose of further analysis we will have to disregard all markers where only an end marker for a seizure exists, as we can only guess when the seizure actually started. Contrary, start markers without a corresponding end marker are of less concern, as we can approximate the duration of the seizure.

Since there can be two start and end markers for the same seizure, they will often overlap one another. For this reason the number of markers does not equal the number of recorded seizures.

Statistic	EEG	Video
Start marker without end	9	2
End marker without start	20	15
Number of cycles with supposed seizure duration over 10 minutes	2	0
Number of cycles with normal duration	82	66

Table 4.3: Statistics for the marker dataset. A normal cycle consists of a pair of start and end markers with a duration under 10 minutes.

The literature suggests that in most cases an epileptic seizure can be identified based on the EEG signal before the video recording [55, 56]. This can be confirmed by looking at our dataset, as in 77 of 93 cases of seizures where both a EEG marker and a video marker are available, the EEG seizure start marker comes before the video seizure start marker.

By analysing the marker dataset further, we can also already define an average time window for seizures in our dataset. This will be important for cases in which no end marker can be found. Figure 4.4 shows a box plot of the duration of the recorded seizures.

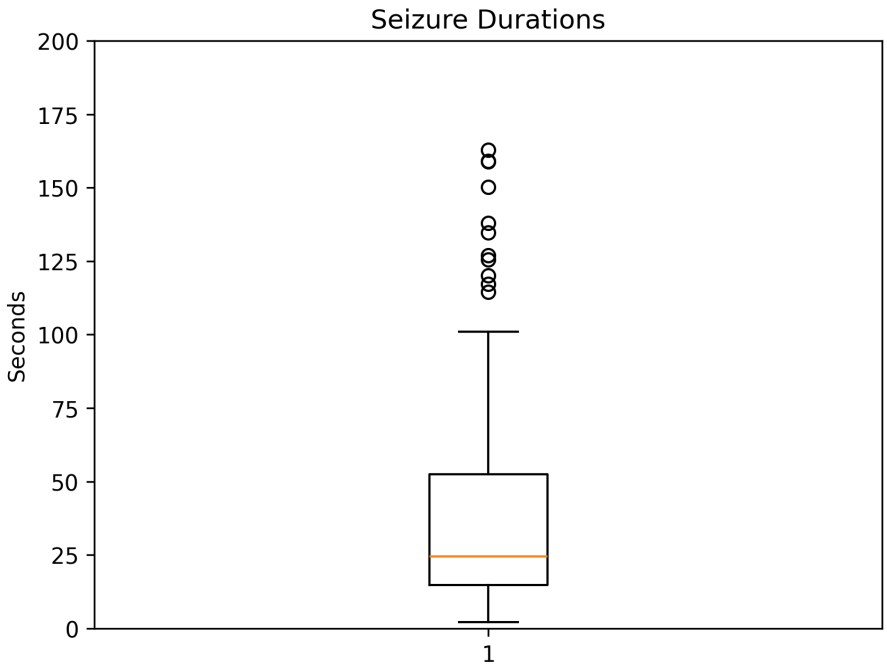


Figure 4.4: Average duration of all seizures with a normal cycle, i.e. start and corresponding end marker. Outliers over 200 seconds not shown.

To see if there is trend of seizures occurrence during a specific time of the day we can take a closer look at the timestamps of the markers. Figure 4.5 shows that there is no trend for patients being more likely to experience a seizure during a certain time of day. There could be some statistical evidence that only certain patients experience seizures more often during a specific time of day, but this is out of scope for this thesis.

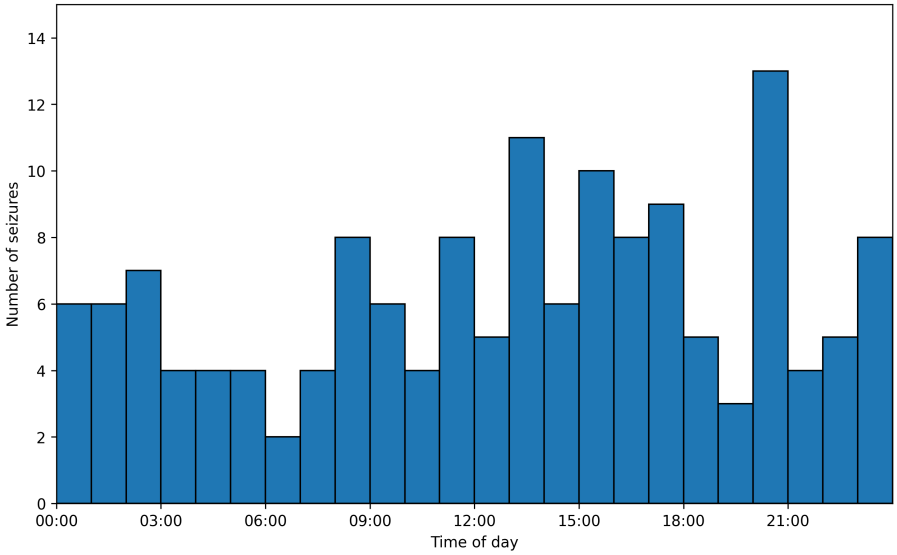


Figure 4.5: Distribution of seizure timestamps.

Lastly, our marker dataset also contains a type column, which gives us all the information about the type of seizure the patient experienced. The different types and their respective counts are displayed in Table 4.4.

Type	Count
Focal Onset - Aware - Motor Onset - tonic	74
Focal Onset - Impaired Awareness - Motor Onset - automatisms	67
Focal Onset - Impaired Awareness - Motor Onset - tonic	50
Focal Onset - Aware - Nonmotor Onset - autonomic	37
Focal Onset - Impaired Awareness - Nonmotor Onset - autonomic	20
Focal Onset - Impaired Awareness - Nonmotor Onset - cognitive	16
Focal Onset - Aware - Nonmotor Onset - behavior arrest	16
Focal Onset - Impaired Awareness - Nonmotor Onset - behavior arrest	15
Focal Onset - Aware - Motor Onset - automatisms	13
Focal Onset - Aware - Motor Onset - clonic	9
Focal Onset - Aware - Motor Onset - myoclonic	8
Focal Onset - Impaired Awareness - Motor Onset - clonic	8
Focal Onset - Aware - Nonmotor Onset - sensory	8
Generalized Onset - Motor - tonic-clonic	4
Unclassified	2
Focal Onset - Impaired Awareness - Motor Onset - myoclonic	2
Focal Onset - Aware - Motor Onset - atonic	2
Generalized Onset - Motor - tonic	1

Table 4.4: Seizure types and their counts.

While the type of seizure was initially assumed to be of more importance, [10] states that both focal and generalized seizures emit the same seizure scent, menthone. Even though it should not be assumed that every type of seizure emits exactly the same VOCs, this should serve as an indicator that an algorithm that can predict one type of seizure might also be used to predict other types of seizures.

Additionally, splitting the dataset according to singular epilepsy variants and trying to predict based on the smaller dataset would be out of the scope for this thesis.

Not all markers currently in the database are of use for our machine learning task. Firstly, we can disregard all end markers since we will not be training on the data during a seizure event, only on inter-ictal and pre-ictal data. The reasoning for this will later be become evident in section 4.3.3. This will already eliminate about half of all markers.

Another pitfall regarding data leakage that we have to watch out for are markers that reference the same seizure. As already explained, there are either one or two start markers for each seizure. This leads to the question which marker should be considered as the true marker for later calculation. When taking a closer look at the non-filtered markers, we can see that professionals usually register a seizure based on the EEG signal first, if both EEG start marker and video start marker are available. There is not any information in our dataset about if a marker belongs to a certain seizure. Meaning that we have to assume that a corresponding pair of start markers (EEG and video start marker) within a conservative 5-minute time window refer to the same seizure event. This is the case for 93 seizure events in our dataset. Of those 93 seizure events, the EEG marker is entered before the video marker in the timeline in 77 of those cases. Meaning that only in 16 cases the video marker comes before the EEG start marker. Most important is the fact that the two marker categories are usually not timed far apart. In the case where the EEG start marker is the first one to arrive, the video seizure start is on average only registered 9.1 seconds later. The other way round, when the video start marker is the first one to be entered, the EEG seizure start event is timed 8.7 seconds later. This shows that choosing which marker to pick as the actual initial starting reference for a seizure event is rather unimportant. For the machine learning task later on, we will always pick the marker which is timed as being the first one of a seizure event, this eases the process and also lets us use every single marker in the dataset. Figure 4.6 also visualizes how seizure events are separated based on the marker dataset.

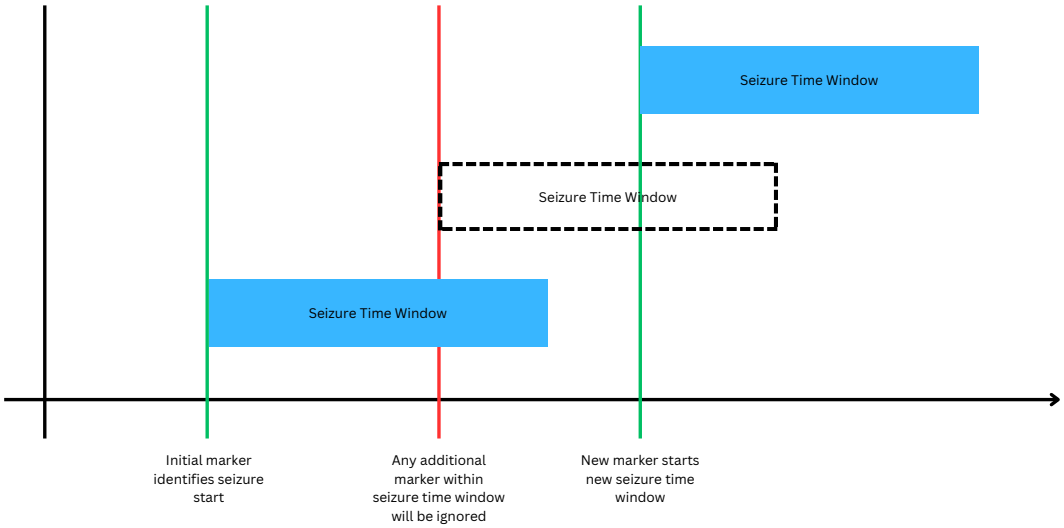


Figure 4.6: Overlapping seizure markers will be ignored.

For future experiments we also make sure that we offer a way to load markers based on the seizure types, e.g. only train on data of seizures with motor onset.

Lastly, we need to filter all marker timestamps at which no VOC data has been recorded, since it is not guaranteed that either a patient was wearing the Smell Inspector sensor at all during their time in the hospital, as they were allowed to opt out of wearing certain sensors, or that the sensors did have enough battery to record and send data.

After going through all these filtering steps we end up with valuable seizure markers from 20 different patients with half of them only having a single seizure during their stay as can be seen in fig. 4.7.

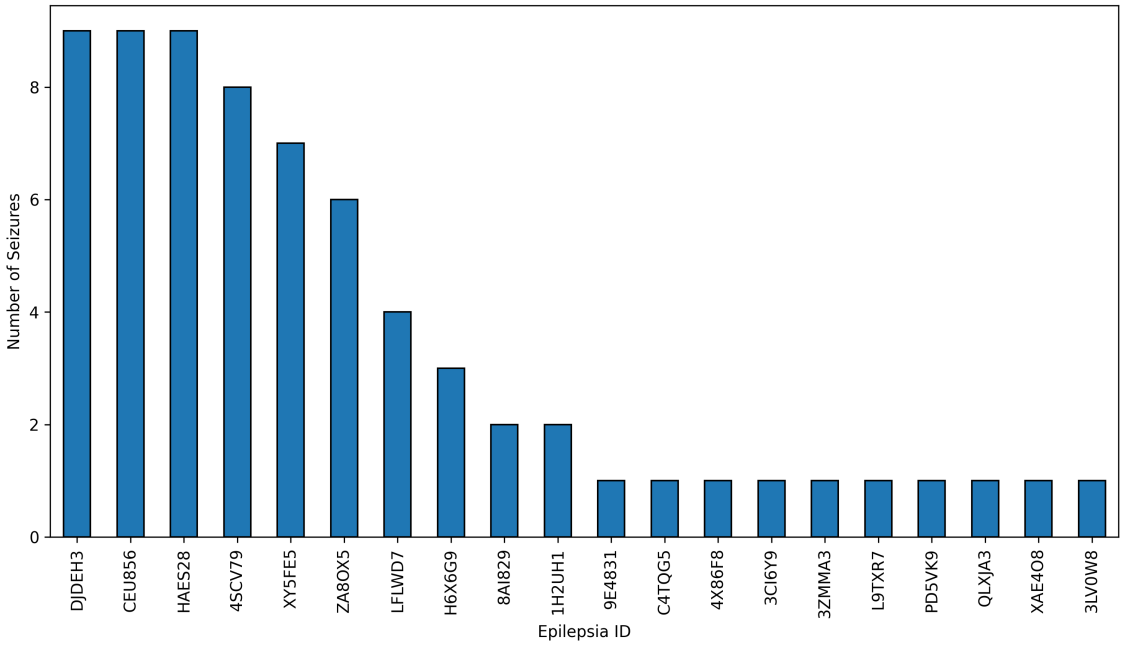


Figure 4.7: Number of seizures per patient after filtering.

In this implementation of the data analysis pipeline a configuration object needs to be provided when loading the ictal markers. This facilitates an easier experimentation phase in which a closer look can be taken on how different settings impact model performance. Table 4.5 shows the options for loading the ictal markers. These options allow filtering and later experimentation with certain seizures or marker types and will be treated as hyperparameters in the grid-search being conducted in section 5.2.

Inter-ictal marker dataset

Almost as important as knowing the timestamps of the seizures themselves, is having a reference baseline with no seizures, so that the distinction can be learned by a model. Since a person with epilepsy is usually not experiencing a seizure, there has been plenty of baseline data collected. The choice of which inter-ictal time windows of patients to pick is not a trivial one and can impact model performance significantly. Choosing these inter-ictal markers should therefore not be done at random. These markers are not manually

Parameter	Description
Exclude start markers	Filter out all start markers (True/False)
Exclude end markers	Filter out all end markers (True/False)
Seizure Types	List of seizure types to include (e.g., `[]`, `['Motor Onset', 'tonic']`)
Marker Type	List of marker types to include (`"eeg"`, `"video"`, `"all"`)
Overlapping allowed	Treat overlapping markers as separate seizures (True/-False)

Table 4.5: Configuration options for ictal marker loading.

created by medical professionals, so an algorithm for obtaining these baseline timestamps has to be created.

The timestamp of an inter-ictal marker should not be too soon after a seizure has occurred, because it is unknown how long seizure scents can linger. Another option is to pick a timestamp leading up to a seizure as an inter-ictal marker. However, as [10] has shown, VOCs can already be emitted one hour before the seizure, so the timestamp before the seizure needs to be picked even earlier. Either way the inter-ictal marker needs to be temporally removed from any other seizure, as to guarantee the best baseline data quality.

While there are only a finite number of epileptic seizures, but virtually endless inter-ictal time windows in the dataset, the proportion of inter-ictal markers to ictal markers in the training dataset, i.e. the label balance, needs to be carefully picked. The topic of label balance will be discussed in section 4.4.3.

Another distinction can be made by separating the origin of an inter-ictal marker. In our study we had 51 people with epilepsy that did not experience a seizure during their time in the hospital. Therefore, in our analysis, the percentage of inter-ictal markers obtained from people with epilepsy that did not experience any seizures in the hospital can be treated as another hyperparameter. Whether introducing data from patients without seizures leads to an increase or decrease in performance needs to be studied. We have to be aware that overfitting could become an issue when deciding against using data from non-seizure patients, but for an initial proof of concept this could still deliver satisfactory results.

Algorithms for obtaining inter-ictal markers

Choosing inter-ictal timestamps for a person with epilepsy with recorded seizures should be done reliably and reproducibly. Therefore, algorithms needs to devised to exactly specify how inter-ictal timestamps are chosen. Since there are two types of patients to pick inter-ictal markers from, patients with and without recorded seizures, two algorithms need to be created. For patients with recorded seizures, their ictal markers serve as a reference point to where the inter-ictal marker should be placed. For patients with no recorded seizures a reproducible approach to choosing a time window needs to be manifested.

As the ratio of inter-ictal markers to ictal markers is an important hyperparameter that can be varied (see section 4.4.3), the algorithms for obtaining inter-ictal markers need to created in way where the number of required inter-ictal markers can be specified at will, whether only 5 are needed or 1000 should not matter to the algorithms. On the one hand this raises the question, which inter-ictal markers should be picked if the number of required markers is low. On the other hand, what happens when the number of inter-ictal markers required is too high, meaning that no more suitable time windows can be found.

Both these problems will be addressed in the following sections.

As is the case with loading the ictal markers in the data analysis pipeline, a configuration object also needs to be provided for loading the inter-ictal markers. Table 4.6 shows the configuration options for loading the inter-ictal markers. It is important to note that the ictal marker loading configuration is also needed for loading the inter-ictal markers, as they are the reference points for creating markers for people with epilepsy with recorded seizures, as will later become apparent.

Parameter	**Description**
Number of markers	Number of markers to be obtained
Percentage without recorded seizures	Percentage of inter-ictal markers coming from patients without recorded seizures
Hour offset for seizure patients	Minimum hour offset from seizure for patients with recorded seizures

Table 4.6: Configuration options for inter-ictal marker loading.

One can notice that in while in table 4.6, the number of markers is required as an argument, in table 4.5 it is not. As the number of seizures markers available is always fixed, there is no need for any argument specifying the amount of markers required.

Patients with recorded seizures

As already mentioned, for patients with recorded seizures, the reference for every inter-ictal marker is always an ictal marker. The important parameter in this case is a temporal offset, defining how many hours the inter-ictal marker should be removed from a seizure marker. This parameter can either be positive or negative, depending on whether the time window should be before or after the seizure. While there is an initial ictal marker that the inter-ictal marker is calculated on, it should be checked whether there is a possible second seizure in proximity to the initial one. If the currently picked time window cannot satisfy the temporal offset for both seizures in the timeline, another time window must be picked.

In the case of a low number of required markers, it cannot be decided which datasets from a participant should take priority over another. It would be possible to simply sort the reference markers, i.e. ictal markers, by either the time they occurred or the EPILEPSIA ID, and pick the first X number of markers. This would however certainly introduce bias into the model. For this reason we will randomly sort the list of ictal markers with a set seed, before looking for suitable inter-ictal time windows. The seed in this instance and all later instance, where randomness will be introduced, will be *42*.

While the initial temporal offset can be entered at will, the case where the number of inter-ictal markers required is greater than the number of ictal reference markers needs to be considered. A fixed temporal offset could at most only match the number of ictal markers at a one to one ratio. Therefore, the temporal offset needs to be increased if not enough time windows have been found yet.

Additionally, in the case of a high number of required markers, a safeguard is implemented to not go into an endless loop. The combination of a high temporal offset parameter with a high number of required markers could lead to a case where too few or no inter-ictal markers can be created.

Patients with no recorded seizures

While there are fewer variables that need to be considered when trying to obtain inter-ictal markers from patients without recorded seizures, a few of the same principles still apply. As already mentioned before, for people with epilepsy that are not in the marker dataset, there is no reference point from which a temporal offset can simply be applied.

As is the case with people with epilepsy with recorded seizures, the possibility of high or low numbers of required markers need to be considered. For a lesser amount of required markers than there are people with epilepsy without recorded seizures the choice will again be made randomly with a set seed of *42*.

Even more difficult than deciding which patients take priority, is deciding which time window of the hospital stay should be picked. As there is no discernible difference in value of the captured data, randomness with a set seed will be the best path forward.

In the case that the number of required markers of this category is greater than the number of people with epilepsy without recorded seizures, the randomly sorted list will be looped through until enough inter-ictal markers have been captured. Each time a new randomly selected time window from the current person with epilepsy is added to the list of inter-ictal markers.

4.3.2 Volatile organic compound dataset

As already briefly touched on in chapter 1 the sensor that has been used to collect data is the Smell Inspector built by Smart Nanotubes. The original casing of the sensor has been adapted to include a microcontroller and an additional battery to ensure longer transmission periods before the sensor needs to be swapped out. The sensor was placed near the armpit on a three point strap to capture as much critical VOC data as possible. Additionally, a fan was built into the adapted casing to guarantee that air is constantly flowing through the sensors. Figure 4.8 shows a picture of how the casing looked like.

Figure 4.8: The Smell Inspector sensor used for VOC data collection.

At its core, the Smell Inspector incorporates four Smell iX16 multichannel gas detector chips, each utilizing fine-tuned carbon nanomaterials as sensor elements. In total the device captures VOC data in 64 different sensor channels.

When air flows over the sensor array, volatile organic compounds (VOCs) interact with the surface of the nanomaterials, either through physical adsorption or weak chemical bonding. These interactions alter the electrical properties of the nanomaterials, such as their resistance or conductivity. Each channel responds differently, depending on its specific surface treatment and the type of gas molecules it encounters. As every sensor channel responds uniquely to different chemical compounds, a distinct signal pattern or "smell fingerprint" for each odour is produced. These signal patterns are then processed using machine learning algorithms to identify and classify the detected odours. [57]

While research previously done with the Smell Inspector showed promising results, showed however that the detector chips may be prone to sensor poisoning over a couple of months. Additionally, experiments to reproduce the same result, showed that scent patterns recorded on different days showed relatively large differences. [58]

For data visualization and analysis, the Smell Inspector can be used in combination with the Smell Annotator software. This software allows users to view and analyse real-time

sensor data, annotate measurements, and store results for further analysis. For this use case we need to work with the raw sensor data, since we do not want any other data processing to interfere with the machine learning model training.

Gases and Volatile Organic Compounds Detectable by the Smell Inspector

Smart Nanotubes provides the following information, seen in table 4.7, on which gases and VOCs are detectable with the Smell Inspector. The categories refer to the possible limit of detection (LOD). An associated value of +2 means that detection of these VOCs is possible even at low concentrations. A value of +1 means that the detectors are well suited for detection and recognition, however only at a sufficient smell intensity. Gases and smells with a value of 0 or even -1 should be avoided and are not detectable.

Table 4.7: Gases and smells detectable by the Smell Inspector.

Gases/Smells	Category	Comment
Ammonia (NH_3), hydrogen sulfide (H_2S), nitrogen monoxide (NO)	+2	Very low LOD (<80 ppb)
Guaiacol, eugenol	+2	Very low LOD (33 ppb)
Phosphine gas (PH_3)	+1	Very low LOD (<10 ppb), condensation of phosphor acid on the detector surface at high ppm concentrations of phosphine
Hydrogen peroxide (H_2O_2), formaldehyde (CH_2O), carbon dioxide (CO_2), ethanol (C_2H_5OH), toluene (C_7H_8)	+1	Reliably detectable, however, at high-ppm concentration levels

Continued on next page

Gases/Smells	Category	Comment
Ketones (acetone, butanone, ...)	+1	Reliably detectable, however, at high-ppm concentration levels
Black tea leaves, green tea leaves, coffee beans, red wine, orange juice, vodka, chocolate, garlic, onion, orange, banana, potato, meat, fish, spoiled banana, spoiled potato, spoiled meat, spoiled fish	+1	Reliably detectable, however, at high smell intensity (like from a closed package)
Water vapor (H_2O)	0	Detectors react on the change of humidity, can be compensated by our software
Hydrogen (H_2), methane	0	Currently not detectable
All neutral gases (nitrogen, argon, helium, etc.)	0	Not detectable
Ozone (O_3)	-1	Can damage sensing material at high concentrations and/or long exposure
Oil and polymer vapors	-1	Can contaminate detectors (no recovery)
Low temperatures, high humidity	-1	Water condensation on detector surface
Very high gas temperatures	-1	Damage to detectors at T>100 °C

Continued on next page

Gases/Smells	Category	Comment
Any liquid	-1	Cannot be used for liquids

As discussed in chapter 3, through analysis of sweat samples collected from people with epilepsy, some VOCs have been identified as possible early warning signs emitted by humans before a seizure, most importantly menthone. As the Smell Inspector works by identifying VOC through their unique fingerprint using different reactive, generic chemical sensors, there is a possibility that menthone is well detectable by the sensor. Even though menthone is not explicitly listed, it belongs in the mentioned group of ketones, because it contains a carbonyl group. If the sensor reacts to common ketones, it would likely react to menthone in the same way at high enough concentrations. Further analysis regarding the chemical similarities between the VOCs detectable by the smell sensor and the identified VOC menthone should be conducted in the future.

Preprocessing the Raw Smell Inspector Data

The schematic in table 4.8 shows the functionality of each channel. The term "raw data" will refer to the actual unprocessed values that were recorded in the hospital and thereafter stored in the InfluxDB. Channels with the same number respond to the same volatile organic compound, meaning there is redundancy built into the sensor. Channels labelled "999" are non-functional base-level channels which should not be taken into consideration when analysing the raw data. In total there 19 non-functional channels and 54 functional ones. Since there are 3 channels for each feature, this leaves us with 15 feature values for each point in time.

Type	Ch1	Ch2	Ch3	Ch4	Ch5	Ch6	Ch7	Ch8	Ch9	Ch10	Ch11	Ch12	Ch13	Ch14	Ch15	Ch16
Type 1	999	999	11	11	11	3	3	3	2	2	2	1	1	1	999	999
Type 2	999	999	999	999	999	9	9	9	6	6	6	5	5	5	999	999
Type 3	999	999	14	14	14	10	10	10	7	7	7	4	4	4	999	999
Type 4	999	999	15	15	15	13	13	13	12	12	12	8	8	8	999	999

Table 4.8: Schematic for all channels of the Smell Inspector.

Analysis of Faulty or Missing Values

For all functional channels the recorded values should always range from zero up to 50000, according to the team at Smart Nanotubes. This means that any recorded values over 50000 should be considered as faulty. Luckily, as there is redundancy built into the sensor, we have multiple options on how to handle faulty values. One option would be to simply take the mean of all three channels for each feature. For this to work it needs to be ensured that the healthy data threshold of 50000 does not get crossed too often or too heavily for it to negatively impact data quality. Figure 4.9 shows how many times this happens for each feature for a single channel given a specific configuration. The configuration used to download this specific dataset can be seen in table 4.9. Some parameters seen in this table will be discussed only at a later point, but are included here for completeness and reproducibility. All statistics that have been computed for this section have been computed based on this configuration. While not completely representative for all different configurations, it does paint a picture of how many data points are usually missing.

Parameter	**Value**	**Parameter**	**Value**
Pre-ictal label percentage	0.5	Exclude ends	True
Inter-ictal label percentage	0.5	Seizure types	[]
Ictal label percentage	0	Marker type	all
Post-ictal label percentage	0	Overlapping allowed	False
Sample time window in seconds	600	Percentage no recorded seizure	0
Pre-ictal definition in seconds	300	Hours offset seizure patients	5
Post-ictal Definition in seconds	600		
Hertz sampling rate	0.5		
Exclude starts	False		

Table 4.9: Configuration with which faulty or missing values have been analysed.

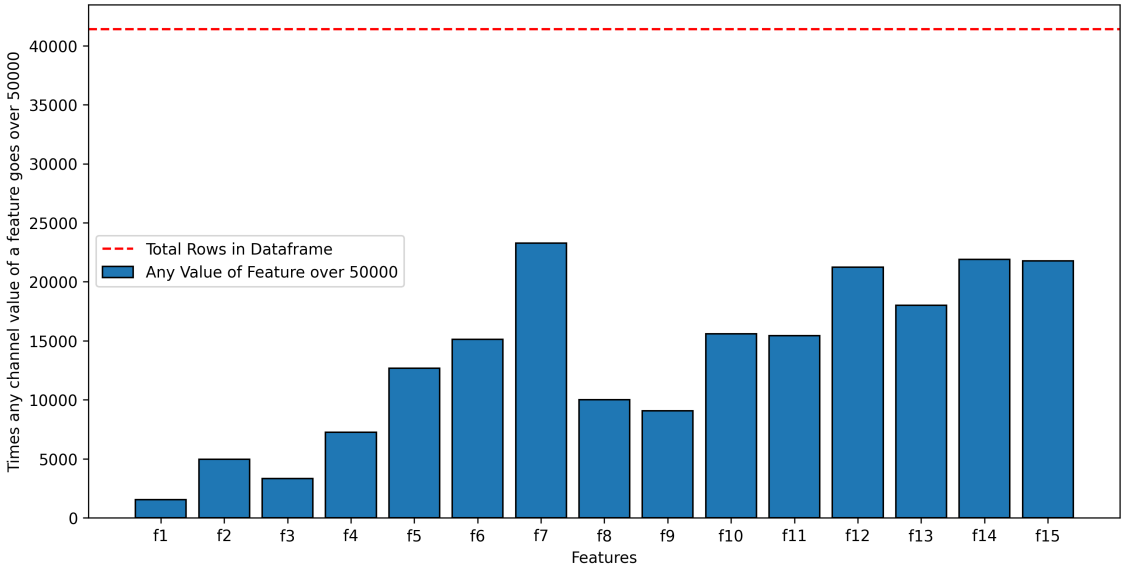


Figure 4.9: Amount of rows where the value of at least one channel of a feature exceeds 50000.

As is evident, it is often the case that a singular channel delivers faulty values. This leads to the question of how much a single faulty channel value would skew the overall mean of a feature. Figure 4.10 shows the average for each functional channel if the channel value is greater than 50000.

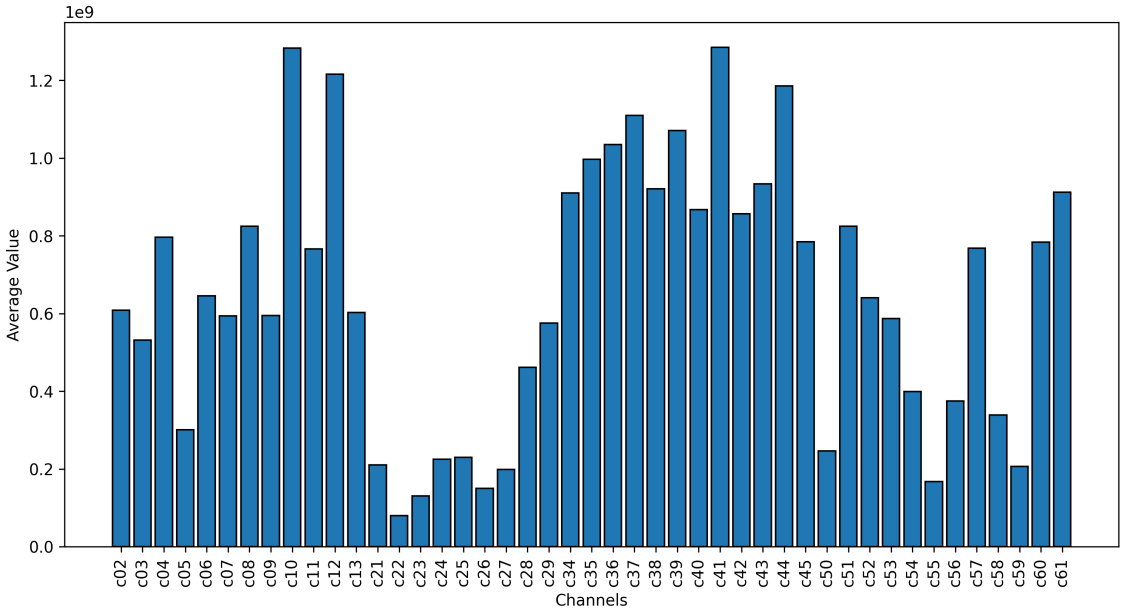


Figure 4.10: Average value for each functional channel when the value goes over 50000.

As can be seen from fig. 4.10 the mean value range for values over 50000 lays approximately between 10 million and 1 billion. This would lead to an extreme skew when taking the mean of three channels, given one channel is not working correctly.

Luckily taking the mean of all channels for a feature is not a necessity. Another approach would be to take the mean of only the healthy values available. As there are three channels for each feature, a single channel with valid data would be enough to work with. However, even with redundant channels included in the sensor, it needs to be checked if in some cases all channels of a certain feature deliver faulty values. Figure 4.11 shows how much data needs to be completely discarded due to the fact that all channels of a feature show values over 50000. Table 4.10 shows the exact percentages where even with redundant channels the recorded data cannot be used.

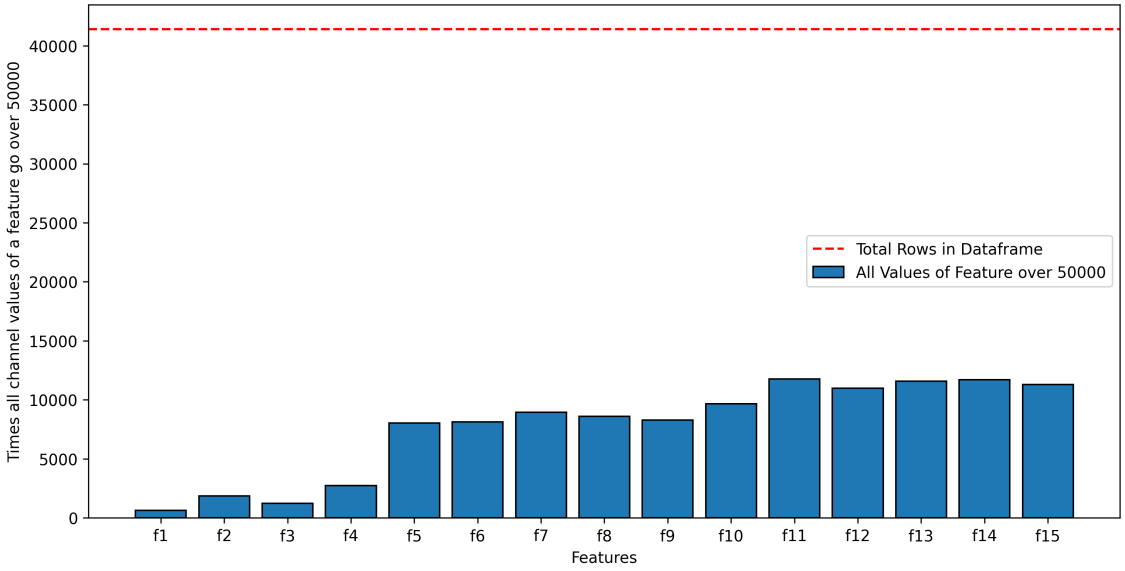


Figure 4.11: Amount of rows where the value of all channels of a feature exceed 50000.

Feature	Value (%)	Feature	Value (%)	Feature	Value (%)	Feature	Value (%)
f1	1.53	f5	19.42	f9	20.03	f13	27.99
f2	4.47	f6	19.66	f10	23.34	f14	28.24
f3	2.98	f7	21.57	f11	28.40	f15	27.26
f4	6.57	f8	20.76	f12	26.54		

Table 4.10: Percentage of unusable data for each feature.

Another statistic that is worth looking at is the distribution of faulty features between positive and negative samples, i.e. ictal and inter-ictal samples, since the pool of negatives samples can be chosen differently if too many features are missing. This is not the case for the positive samples, since the ictal markers cannot be readjusted or other time windows picked. Figure 4.12 shows the distribution of positive and negative samples for all cases where all values of a feature are over the healthy threshold.

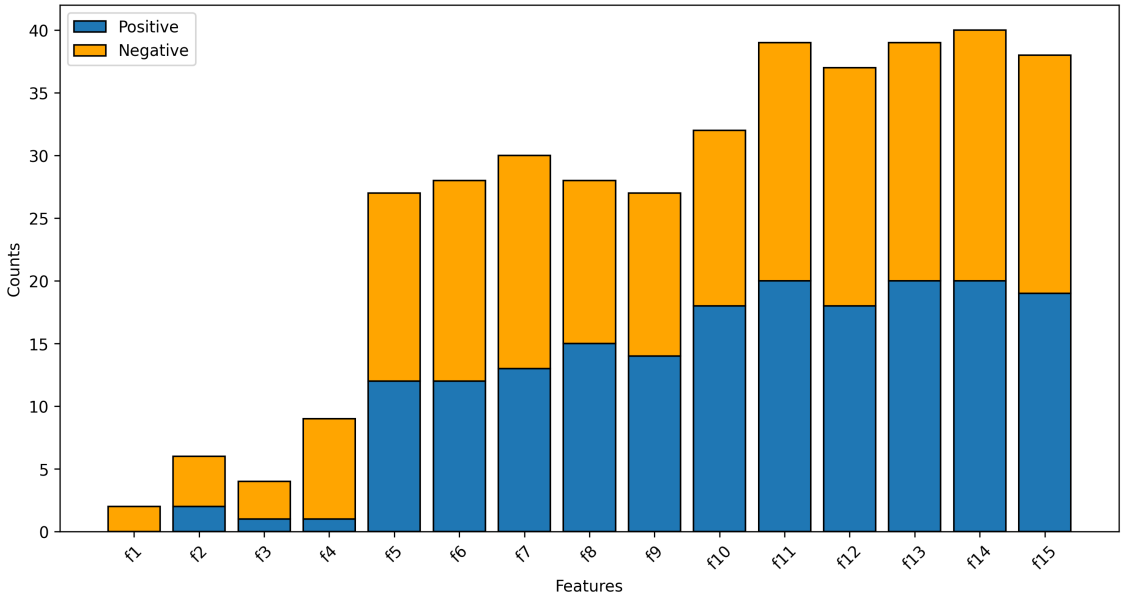


Figure 4.12: Distribution of faulty data between positive and negative samples for completely missing features.

Although some features have nearly 30 percent unusable data, it is important to analyse the distribution of faulty data at the smell sample level, as the overall average does not provide the full picture. If a significant portion or all features of a single smell sample (e.g., a 5-minute pre-ictal time window) are unusable, the sample must be excluded from the training data. However, this exclusion is not necessary if the average applies uniformly across all smell samples, as 70% healthy data would still be sufficient for our purposes. Figure 4.13 illustrates the percentage of completely missing features for each sample, sorted in decreasing order.

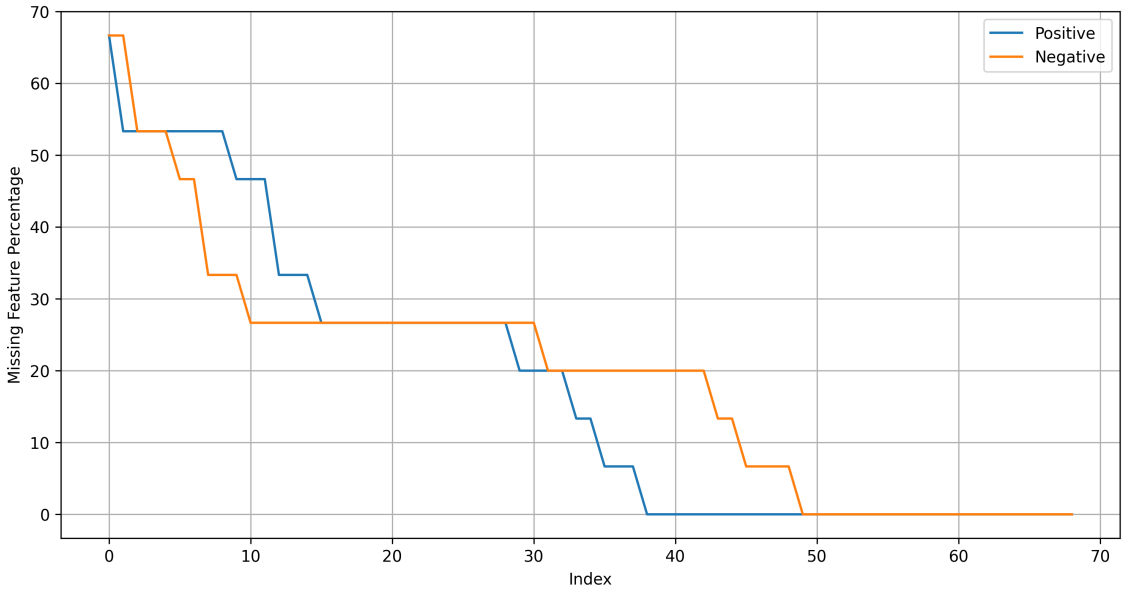


Figure 4.13: Percentage of missing features for each sample in decreasing order.

As can be seen, there are only some instances in which the number of completely missing features goes over 50%, which is not ideal. However, as we are already dealing with a very low number of positive samples, we are not going to filter the training and test set any further.

Handling Faulty or Missing Values

Whether features are completely faulty or only a few seconds worth of values are missing from a 10-minute time window, the values need to be imputed in one way or another. According to recent research there are two possibilities on how to handle faulty or missing data. The first option is to use statistical imputation to correct faulty values and fill in missing values. Imputation has been shown to result in good performance when the missing or faulty data is hiding useful information. These performance increases have been shown when working with categorical data [59]. It cannot be said with certainty that the missing and faulty values hide any useful information in our case. Additionally, for all cases where a feature is missing completely, i.e. there is not a single reference value,

from a time window, all imputation methods will have to rely on data from other samples, which can lead to lower data quality.

The other way to handle faulty or missing data is to explicitly denote in the dataset that a certain value or a certain feature is missing or faulty. This can be done using one-hot encoding, meaning that each feature gets a corresponding binary column which will contain the value "1" if a correct value is present or a "0" if the value should be ignored. However, this could lead to the model learning patterns about the missingness of certain features, which would be counterproductive, since we are dealing with a MCAR (Missing Completely At Random) problem. Research has shown that encoding missingness can be helpful in cases where the missing data rate is high [60].

For our dataset there are two possibilities we will try out in the grid-search. Firstly, we will try a statistical imputation approach to deal with missing data. In the case of a partially missing feature, meaning there is at least a singular reference value in the time window, a simple forward fill followed by a backward fill will be used to fill in the missing values or overwrite the faulty values. More complex and accurate approaches could be taken if the experimental analysis shows promise, but this is out of the scope for this thesis. In all cases where there is not a single value that can be used for the forward and backward fill, the mean of all values, including other samples, for that feature will be used. This is a safer approach than simply capping the values at 50000, since this could skew the data.

With the second approach that will be tried we will replace all faulty values with the value negative one. The reasoning behind this being that with enough data the model could infer and learn that -1 means that this value can safely be ignored. It cannot be said with certainty which one of these approaches is better suited for this use-case, which is why they will both be tried out in the grid-search.

Additional metrics

Since our marker dataset is relatively small we need to incorporate as many data points for each seizure so that a machine learning model can still function effectively. For each of the 15 feature channels we calculate the standard deviation, variance, mean, maximum and minimum over a 20 timestamp range. The timestamp range depends on the sampling rate in which we load the data into a model. For example, if we load the data at a 0.2 Hz

sampling rate, meaning one row of data each 5 seconds, the range over which the metrics are calculated is 100 seconds. This hyperparameter optimization will further be elaborated on in section 4.4.2.

This results in a maximum of 97 columns of values for each recorded time step for a single patient, including time of day information alongside features and metrics. Deciding which metrics to include in the model is treated as a hyperparameter. While standard deviation and variance are considered essential features to include in the dataset, adding maximum, minimum, and mean features for each channel may negatively impact results if the data is overly skewed. Therefore, during the grid-search, models will be trained both with and without the maximum, minimum, and mean features to evaluate their impact.

Normalization of Values

Working with a normalized set of values can improve machine learning performance [61], which is the reason why we will apply Z-score normalization (see eq. (4.1)) on a global and not patient-specific level. As there have already been additional metrics calculated, we are working with a hybrid approach to capture information on both a global and patient-specific level. The Z-score of a value indicates how many standard deviations the value is from the mean of the dataset. The formula for calculating the Z-score of a value x is given by:

$$z = \frac{x - \mu}{\sigma} \tag{4.1}$$

where:

- x is the value to be normalized,
- μ is the mean of the dataset,
- σ is the standard deviation of the dataset.

By applying Z-score normalization, the transformed data will have a mean of 0 and a standard deviation of 1.

Removing outliers

Per default, our processing pipeline for the grid-search will remove any outliers before calculating any other feature, to ensure that any skew in the data remains minimal. This happens even after capping the data according to the healthy data threshold, which for our grid-search will be 50000. The threshold might not remove all outliers, which is why the top two percent of each channel will be disregarded and later calculated again according to the missing value logic described in section 4.3.2.

4.3.3 Data structure

All the following analysis and model-based classification has been done using SKTime [62], which provides a unified interface for time series machine learning tasks. Additionally, SKTime is equipped to handle multivariate time series data, where multiple variables are observed over time. Training a model using SKTime requires the data to be in a specific format. In the following the structure and format of input and target values are going to be discussed.

Input values

As many other machine learning frameworks, SKTime uses the data-container library Pandas [63] under the hood. For this reason, the SKTime requires the input data to be provided in a Pandas multi-index. There are other options, such as a NumPy [64] 3D array, but these come with small drawbacks as per the documentation.

The multi-index in our case has two levels. The first will represent a sample level, corresponding to a specific continuous time window, e.g. 5 minutes, of a singular patient. There can be multiple samples of the same study participant at the first level. For example, for model training purposes we could use one pre-ictal sample and one post-ictal sample of the same person with epilepsy. On the second level we have the time index containing timestamps, which will always be roughly equidistant to one another. It is important to note, that for each sample the number of timestamps provided needs to be same. To simplify the data, the timestamp will only contain the number of seconds passed in that particular day, as machine learning frameworks typically cannot work with non-numerical

input data. It is not necessary to denote the timestamp in milliseconds as the smell sensor recorded data in the 0.55 hertz range. Correctly specifying the time of day can be of value, as SKTime also considers the value of the time index as a feature to learn from. The actual date on the other hand can be disregarded, as nothing can and should be learned from the date itself. The time index can be arbitrarily long and granular, meaning later we will try to optimize between different time windows and aggregate values. Table 4.11 shows an example of how the training data would look like. The real training data of course has between 60 and 97 columns depending on which features should be included in the model. Within these columns the raw VOC data and calculated features based on the raw data, such as variance and rolling averages, are contained.

Sample index	Seconds of day	Feature 1	Feature 2	Additional Feature
0	42331	41051	12566	1.385
	42333	41055	12570	1.322
	42335	41062	12568	1.334
	⋮			
	42627	41231	12407	1.385
1	3012	20392	31828	-0.021
	3014	20385	31799	-0.023
	3016	20380	31789	-0.025
⋮				

Table 4.11: Structure of the pandas multi-index as input data.

In the case of missing values for a specific time step, the techniques described in section 4.3.2 will be used. Models cannot be fitted onto input values with non-numerical values, so it is necessary to ensure every row and column contain a value.

Target values

As we are working with a supervised approach, we need target values in addition to the input data, from which a model can learn. These target values will correspond to the sample index of the training data and therefore label the sample correctly. There are two

options on how the training data can be labelled. Firstly, a differentiation could be made between four states:

- pre-ictal
- ictal
- post-ictal
- inter-ictal

or simply between two states:

- pre-ictal
- non-pre-ictal

For our use case it is more effective to simply train the model on the binary label, only distinguishing between pre-ictal and non-pre-ictal states. Further reason on why a binary classification suffices, is the lesser importance of being able to tell apart the other states. Having the current end goal in mind, only the correct prediction, i.e. the pre-ictal state, would make a meaningful difference.

Of course from a research perspective, it would be of interest if a model could be trained that can distinguish between each of the states. Although, the problem could arise that a combination of VOCs, that the model learned are predictors for a seizure, could leak into the other phases. Meaning that the critical VOC combination could still be in the air during the post-ictal state, making a distinction between pre- and post-ictal states difficult.

Table 4.12 shows the simple array of target values that is subsequently provided to the model.

Sample index	Label
0	pre-ictal
1	pre-ictal
2	non-pre-ictal
⋮	
263	non-pre-ictal

Table 4.12: Structure of the target values dataframe.

Meta-data structures

For later data split and analysis purposes we use a third data structure that gives some additional information about the samples. Table 4.13 shows a few example rows of the information we are storing. There is reason to suspect that certain data splits containing only EEG markers or only a certain type of seizure might result in a better performing model, which is why this meta information is included in further analysis. This array does not however get passed into the framework, i.e. this is not information that is going to be learnt from.

Sample index	Epilepsia ID	Type	Marker
0	XY5FE5	Focal Onset - Aware - Nonmotor Onset - autonomic	EEG
1	XY5FE5	Focal Onset - Aware - Nonmotor Onset - autonomic	EEG
2	CEU856	Focal Onset - Aware - Motor Onset - tonic	Video
⋮			
263	NV46I0	Focal Onset - Impaired Awareness - Motor Onset - automatisms	EEG

Table 4.13: Dataframe storing the meta-data of samples.

4.4 Hyperparameter

As already touched on in some previous chapters, when training a model to accomplish the task of predicting epileptic seizures, there are a lot of possible hyperparameters that can be adjusted to improve the resulting performance. Finding well performing training data related hyperparameters, i.e. what information is being feed into a model, is in our case even more impactful than tweaking model-specific hyperparameters.

Model-specific hyperparameters will not be discussed in this chapter, but rather in their respective sections in section 5.2.

Many hyperparameters have already been introduced in section 4.3. To give a final overview over all possible hyperparameter, including new ones from this chapter, a summary will be shown in section 4.4.4.

Figure 4.14 shows which classical hyperparameters need to be determined before a grid search can be conducted.

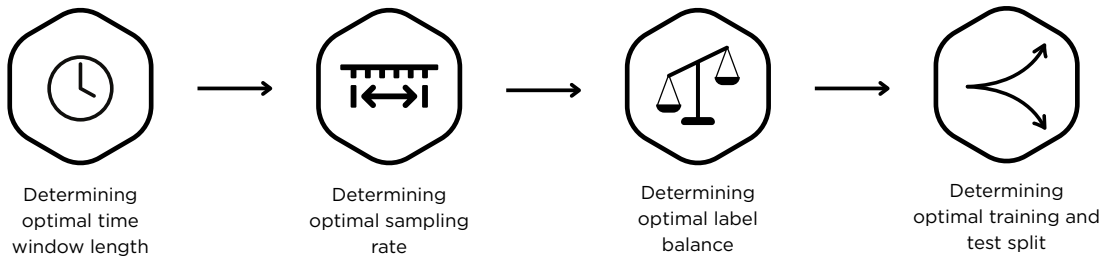


Figure 4.14: Overview of the steps to determine hyperparameters within the pipeline.

4.4.1 Time window length

Optimal time window length for prediction of epileptic seizures is a well researched subject, even though most research has been done using EEG signals, not with VOC data. There is a consensus that the optimal prediction window for pre-ictal states lies between 10 and 60 minutes. While [65] states that the optimal pre-ictal time window varies greatly from patient to patient, this optimization is out of scope for this thesis. For most cases a

time window of 30 minutes was deemed as the best option for training machine learning models [66, 67].

One important pitfall one has to avoid when trying to predict seizures is label poisoning, meaning that we strictly want to classify the time before the seizure as pre-ictal. This means that any overlap between these periods and the ictal period should be avoided to prevent label poisoning. Therefore, a buffer period should be introduced before each seizure marker to ensure that the pre-ictal markers are not contaminated by ictal states. For our purpose we introduce a buffer period of X seconds before each marker to strictly separate pre-ictal and ictal data. This X can be chosen at will in the grid-search, since the exact amount of buffer time needed is not known. We will vary the buffer time between 2, 5, 10 and 30 minutes to assess which models trained with different buffer times achieve the best results.

4.4.2 Sampling rate

Sampling rate is in our case defined as the time between recorded values. It is inherently limited by the sampling rate of the Smell Inspector itself, which is about 0.55 Hz or 1.8 seconds between recorded values. Any value higher than 0.55 Hz would provide no extra information when downloading the data from the Influx DB. A lower sampling rate would mean that we can feed longer time windows into a machine learning model without actually using every single value but averaging values over a certain time period.

There is some evidence that lower sampling rates suffice for many machine learning tasks. This research was concerned with classifying physical activity and detecting falls, using an accelerometer with sampling rates between 25 Hz - 100 Hz. While a higher sampling rate was used in this case, the analysed time window was consequently shorter [68, 69]. The same trend could possibly be observed for our use case, where lower sampling rates would achieve better or equal results if the analysed time window is long enough. The only problem with this assumption would be that it is not certain whether just stretching out the analysed time window would also lead to better results. If the critical changes in the VOC data only occur a few minutes before the start of the seizure, enlarging the time window would only include more noise into the analysed sample.

For this reason, as we are already on the lower end of resolution for machine learning tasks, a sampling rate of 0.5 Hz will be used, meaning one row of values every two seconds, therefore slightly undersampling the original data, but essentially not aggregating any data points together.

4.4.3 Label balance

Ensuring a balanced ratio of positive to negative labels is crucial for achieving good model accuracy. A 1:1 ratio is generally considered ideal for our evaluation, as it minimizes bias towards the larger class and promotes a more generalized model that performs well on both classes [70]. However, there are exceptions to this rule. For instance, random forest models can sometimes perform better with a higher ratio, such as 1:10 [71]. While techniques like over- and undersampling can be used to balance datasets, we have sufficient positive and negative labels available, making it unnecessary to use such methods.

For our further evaluation we will go with the conservative approach and try to keep it as close to a 1:1 ratio of positive and negative labels as possible. What will need to be investigated when experimenting with model evaluation later on, is how strongly the choice of which negative labels are picked, affects the models' performance. For positive labels, there is no other choice than to work with the markers that have been given to us by experienced doctors (disregarding smaller hyperparameters, such as choosing the exact time window based on the ictal marker). As already discussed in section 4.3.1, there are more inter-ictal markers available than ictal markers. This leads to the question, which inter-ictal markers should be used when the label ratio should not be skewed too much. One approach might be that for every ictal marker, i.e. every positive example, a complementary negative example from the same patient is used. If the performance shows promise using this approach, it would be the ideal scenario as it would showcase a clear difference between the pre-ictal state and the inter-ictal state in the VOC data. Using this approach, it needs to be tested whether it is better to pick an inter-ictal marker leading up to the seizure, e.g. 3 hours before, or some time after the seizure has ended.

Other approaches might include inter-ictal markers from patients with no recorded seizures. This could however lead to the model becoming biased towards learning that some patients never have positive labels.

Since our grid-search will be conducted using a time-series specific CNN, a 1:1 ration is ideal and will be used in the search. No other configurations regarding label balance will be tried out, since varying other hyperparameters instead seems more promising.

4.4.4 Summary of all possible hyperparameter

Being at the end of this chapter, all hyperparameter and their possible values seen in section 4.4.4 have now been thoroughly discussed.

Hyperparameter	**Data Type**
Marker Type	List of Strings
Seizure Types	List of Strings
Overlapping Markers Allowed	Boolean
Exclude Seizure Start Markers	Boolean
Exclude Seizure End Markers	Boolean
No-Seizure Data Percentage	Float
Hour Offset For Inter-Ictal Markers	Integer
Pre-Ictal Label Percentage	Float
Inter-Ictal Label Percentage	Float
Ictal Label Percentage	Float
Post-Ictal Label Percentage	Float
Sample Time Window	Integer
Pre-Ictal Definition	Integer
Post-Ictal Definition	Integer
Sampling Rate	Float
Faulty Data Value Cut-off	Integer
Add Normalized Features	Boolean
Remove Raw Features	Boolean
Additional Features	List of Strings
Missing Value Strategy	String
Add Time of Day Feature	Boolean
Remove Top Percentile (Outliers)	Integer
Train-Test Strategy	String

Table 4.14: All hyperparameters that will be considered in the grid-search.

Marker and seizure types refer to the corresponding strings we received from the medical staff. Overlapping markers and the impact of allowing them has been explained in section 4.3.1. Excluding seizure start and end markers have been added as options, since

depending on the experiment one only wants to work with a certain type of marker. *No-Seizure Data Percentage* refers to the amount of inter-ictal samples from patients with no recorded seizures that should be included in the input data. The rest of inter-ictal samples will come from patients who experienced a seizure in the hospital. The hour offset describes the minimum amount of time that should have passed since the last seizure for the inter-ictal marker to be valid. The label percentages in summation need to be equal to 1. Sample time window and pre-ictal definition as well as sampling rate have been explained in the sections above. Faulty data value cut-off describes the limit where a value from the database is still considered healthy. The missing value strategy describes the way missing or faulty values should be handled with the options being *mean* and *negative-one* as per section 4.3.2.

Additional features that should be calculated based on the raw features should be provided in an array of strings. Even though more training data and features is usually welcomed, the removal of the raw features captured by the sensor is treated as a hyperparameter. It is not clear whether including the raw values in the training data might not mislead the model during training. The same goes for adding the normalized values of each feature. Since there is no literature on our specific use case, each combination of these two hyperparameter will be tried out during the grid-search.

The time of day feature will always be included in the training data, since we see no reason why any variation would make sense for the grid-search. The same is the case for the top percentile removal, which will be fixed at two percent.

Now it is time to discuss the last hyperparameter, which is also already included in section 4.4.4, the training and test split.

4.5 Training & test split

There still remains the question on what part of the input data the training should happen and which part should serve as a testing set of samples. In this chapter, we will discuss the methodology used to split the dataset into training and testing sets, ensuring that every model is evaluated on unseen data to provide an unbiased estimate of its performance.

There are multiple ways to split up data, each with their own positive and negative effects. The aspect of how the model is going to possibly used in the future also needs to be considered. It is clear that a random training test split would lead to data leakage, which is why this method will not be discussed or considered.

4.5.1 Leave one patient out (LOPO)

Leaving one or multiple patients out of the dataset is beneficial if the generalization of a model should be prioritized. In fact, a great portion of recent research on the topic has been done using the LOPO strategy [72, 73]. While performance might take a hit, LOPO is the only way to determine whether a completely unseen patient can rely on the accuracy of the model. In the context of EEG based seizure prediction Shafiezadeh et al. used a patient-based split to show that cross-validation does not automatically lead to good generalization of a prediction model [74]. However, the performance of LOPO can vary significantly depending on the patient or patients held out. Some patients might have more complex seizure data, leading to less consistent results.

The problem we see in our case with LOPO is that preliminary results already showed that the task at hand is a difficult one. Our goal is to simply check the feasibility of working with VOC data when it comes to epileptic seizures. Achieving good results with the LOPO strategy would go beyond what we are trying to achieve and is therefore out of the scope for this thesis.

4.5.2 Temporal split

Temporally splitting up the data means using the last seizure of a patient with multiple seizures as a testing seizure on which the model performance can be evaluated. While [75] discusses a different topic, it is clearly being stated that a temporal split is common for time series machine learning tasks. One major advantage of a temporal split would be that it mimics real-world scenarios where past seizure information could be used to predict future ones. It also respects the chronological order of data, which is particularly important for time-sensitive patterns, as it avoids any future data leaking into the training process. Temporal splits can capture evolving trends in seizure patterns within a patient,

especially if there are gradual changes over time due to factors like medication or other adjustments.

However, temporal splits have their drawbacks. One issue is that the training data might not include enough variability if the earlier data does not fully represent the range of seizure patterns the patient experiences. If the distribution of seizures changes significantly over time, the model trained on older data may struggle to generalize to new patterns in the test set. Temporal splits also require sufficient data across time to work effectively. Only seizures of patients with multiple recorded seizures could be picked as part of the training set. While half of patients did experience more than one seizure if they experienced any at all, this issue needs to be addressed. For training purposes the data from patients who only experienced a single seizure can be used, but these seizures cannot be in the test set, since we want to ensure that every patient has been seen before by the model.

In summary, this approach is less useful for generalizing across patients, as it focuses on within-patient predictions. As we have to factor in that there is already a high variance and bias in our current dataset using a temporal split would ease the task at hand. It is not our goal to build a fully robust machine learning model that can generalize across patients without problems. For that reason, in the evaluation later, we will train on the last seizures of patients. To be exact, the last 30% of seizures (rounded to the lower whole number) will be used for the testing set.

5 Results

As with all other chapters fig. 5.1 shows the final steps that need to be taken to finally arrive at a result. Firstly the right performance metrics for the task need to be determined. With these performance metrics in mind the grid search and therefore all 576 models can finally be trained and evaluated. Thereafter, the aggregated results and additionally the results of the best performing models will be highlighted. Finally, as we need to make sure the performance of the best model is not just based on spurious correlation, extensive robustness checks of the model will be performed.

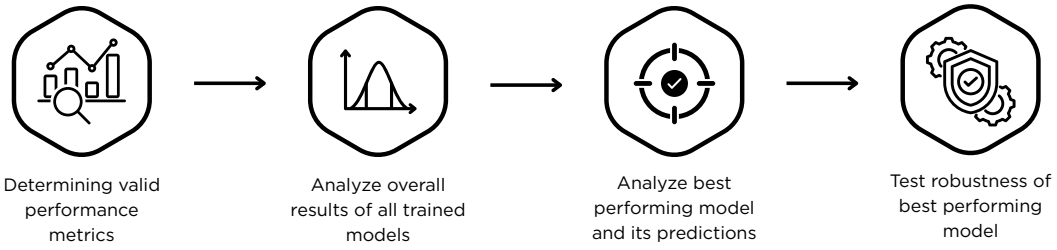


Figure 5.1: Overview of the evaluation steps within the pipeline.

5.1 Performance metrics

As we will be evaluating a multitude of models with different hyperparameters, we need a way to assess these models. While the list of possible performance metrics is endless, we have to make a consideration which metrics are most suitable in this context [76].

When trying to predict epileptic seizures a low false negative rate is of utmost importance. Missing an imminent seizure could have fatal consequences for people with epilepsy. For this reason, especially metrics that capture the rate of false negatives will be included in our comparisons.

Following are the performance metrics that will be used and their definitions. The literature accepts these metrics as generally useful when evaluating models, especially in a medical context [77, 78]. While some metrics may be more important than others for this use case, neglecting others can lead to overall worse practical results, e.g. recall of 1 could also mean that everything has been classified as a pre-ictal phase.

For all following formulas the following abbreviations are used.

- True Positives (TP): The number of positive samples correctly classified as positive.
- False Positive (FP): The number of positive samples incorrectly classified as positive.
- True Negatives (TN): The number of positive samples correctly classified as negative.
- False Negatives (FN): The number of positive samples incorrectly classified as negative.

5.1.1 Accuracy

Accuracy is a metric that measures the proportion of correctly classified samples (both positive and negative) out of the total number of samples. Accuracy can be calculated using the following formula.

$$\text{Accuracy} = \frac{\text{True Positives (TP)} + \text{True Negatives (TN)}}{\text{Total Number of Samples}} \tag{5.1}$$

5.1.2 Recall

Recall, also known as sensitivity or the true positive rate, is defined as the proportion of actual positive samples that are correctly identified by the model. Recall can be calculated using the following formula.

$$\text{Recall} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Negatives (FN)}} \tag{5.2}$$

5.1.3 Precision

Precision, also known as Positive Predictive Value, is defined as the proportion of predicted positive samples that are correctly identified as positive by the model. Precision can be calculated using the following formula.

$$\text{Precision} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Positives (FP)}} \tag{5.3}$$

5.1.4 F1 Score

The F1 Score is the harmonic mean of Precision and Recall, providing a single metric that balances both the Precision and Recall of a model. It is particularly useful when the dataset is imbalanced. The F1 Score can be calculated using the following formula.

$$\text{F1 Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{5.4}$$

5.1.5 ROC AUC

The Receiver Operating Characteristic - Area Under the Curve (ROC AUC) is a metric that evaluates the ability of a model to distinguish between classes at different thresholds. It is calculated as the area under the Receiver Operating Characteristic (ROC) curve, which plots the True Positive Rate (Recall) against the False Positive Rate (FPR) at various threshold settings.

The ROC AUC score ranges from 0 to 1, with a value of 1 indicating perfect performance, 0.5 indicating no differentiation (equivalent to random guessing), and values below 0.5 indicating worse-than-random performance.

$$\text{ROC AUC} = \int_0^1 \text{TPR}(t)\, d\text{FPR}(t) \tag{5.5}$$

Where:

- $\text{TPR}(t)$: The True Positive Rate (Recall) at threshold t, calculated as $\frac{\text{TP}}{\text{TP+FN}}$.
- $\text{FPR}(t)$: The False Positive Rate at threshold t, calculated as $\frac{\text{FP}}{\text{FP+TN}}$.
- t: The decision threshold used to classify positive and negative samples.

5.2 Evaluation

Now that all necessary hyperparameters have been defined and thoroughly discussed, we can start looking at how differently configured models perform with the data that has been captured and preprocessed in the EPILEPSIA study. In preliminary experiments it has become evident that the most impactful factors are the data-related hyperparameters and not any model-specific hyperparameters. In addition, these experiments showed that it is not easy to find optimal hyperparameters that result in convincing performance of a model. To maintain reproducibility and not miss any well performing configuration of hyperparameters a custom implementation of a grid search algorithm will be used. After taking a closer look at the results of the grid search, the impact of the varied hyperparameters will be assessed.

5.2.1 Parameter Grid

After having thoroughly dissected each and every hyperparameter that could affect the ability to predict seizures beforehand, it is now time to select ranges in which a grid search should be performed. This simple approach enables us to try out each and every combination of hyperparameters, to see where the potential of seizure prediction lies. The selection of value ranges needs to be handled carefully as too many combinations will lead to an unreasonably long model training time. Table 5.1 shows each option that will be tested with the given model. The number of possible combinations for this parameter grid is 576, i.e. 576 differently configured CNN models will be trained and tested. Having prioritized the more impactful hyperparameters, this is right on the edge of acceptable runtime while also exploring as many options as possible.

The focus of our grid-search laid heavily on the exact timing of the time windows that are to be picked, since we saw the most potential in this category of hyperparameters. Previous chapters should provide the detailed reasoning to why certain value ranges were chosen. The most important decisions will be highlighted here again.

We made no discrimination between different seizure types or the origin of the marker. While it may be possible to correctly predict seizures easier for some seizure types, we wanted to test if a generalized approach is possible.

No inter-ictal markers from patients without any recorded seizures were included in the grid search, in preliminary experiments it has shown that the VOC profiles of patients differ heavily, to the point where the model learned the differences in the VOC profiles themselves and not the temporary seizure.

The choice which time window to pick for inter-ictal markers, originating only from patients where there is also seizure data available, has been limited to either 5 hours before, 5 hours after or 10 hours after the seizure. In further experiments this time range could be increased even further, away from any lingering seizure scents.

As already explained we will limit the grid search to training on binary labels, no data from ictal or post-ictal phases will be included in the training.

With the variation of sample time windows and also offset from the seizure itself (pre-ictal definition), we wanted to test the impact different time windows have on the results.

Parameter	Values
Marker Type	All
Seizure Types	All
Overlapping Markers Allowed	No
Exclude Seizure Start Markers	No
Exclude Seizure End Markers	Yes
No-Seizure Data Percentage	0%
Hour Offset For Inter-Ictal Markers	{-5, 5, 10} hours
Pre-Ictal Label Percentage	50%
Inter-Ictal Label Percentage	50%
Ictal Label Percentage	0%
Post-Ictal Label Percentage	0%
Sample Time Window	{5 min, 10 min, 20 min}
Pre-Ictal Definition	{2 min, 5 min, 10 min, 30 min}
Post-Ictal Definition	10 min
Sampling Rate	0.5 Hz
Faulty Data Value Cut-off	50,000
Add Normalized Features	{Yes, No}
Remove Raw Features	{Yes, No}
Additional Features	{[Standard Deviation, Variance], [Standard Deviation, Variance, Mean, Max, Min]}
Missing Value Strategy	{Negative One, Mean Imputation}
Add Time of Day Feature	Yes
Remove Top Percentile (Outliers)	2%
Train-Test Strategy	Last Seizures

Table 5.1: Grid-search parameters. Values in curly brackets have been varied in the grid-search.

We wanted to check whether including normalized features might help the model learn, which is why adding normalized features as well as removing raw features will be tested in the grid search. For the additional features that are added in each model test, we always include the standard deviation and variance, but also try adding the mean, minimum and maximum in half the tests.

For the missing value strategy taking the mean value imputation of a feature as well as the negative one strategy will be tried out.

Lastly, we vary the trained models by trying out the random train and test split as well as the custom split which is described in section 4.5, where we train on the first 70% of

seizures and test on the last 30%. This surely simplifies the testing for all models, but as the goal of this grid-search is a preliminary proof of concept, we think it is appropriate to split up training and testing data like this. In addition, a product capable of detecting seizures beforehand could also behave similarly. It would also learn from all previous experienced seizures, in addition to the ones already learned during training, and try to predict on all upcoming seizures.

5.2.2 Choice of classifier

We are limiting the scope of this research to the functionality of the SKTime library. Specifically their implementation of the CNN classifier described in [14]. While SKTime offers multiple classifiers that can handle multivariate data, like we have in our case, we wanted to focus this work on the grid-search and optimization of hyperparameters. The reason for selecting the CNN classifier is its ability to handle multivariate time series data effectively and its proven performance in similar tasks. The architecture of CNNs allows them to capture spatial and temporal dependencies in the data, which is crucial for accurate seizure prediction. Additionally, the implementation in the SKTime library is well-documented, provides a user-friendly interface and is optimized for performance, making it a suitable choice for our experiments. Models like LSTM and transformer models are more complex and resource-intensive, which is why they were not chosen for this analysis. Given the preliminary nature of our proof of concept, we opted for a model that balances performance with computational efficiency.

The Convolutional Neural Network (CNN) classifier is configured with several default settings that influence its training and performance, which have not been changed for the grid-search. The setting that has been used for the grid-search is written in the bracket.

- Number of Epochs (2000): The number of epochs for which the model will be trained. A single epoch is one complete pass through the entire training dataset. More epochs can lead to better results but may also increase the risk of overfitting.
- Batch Size (16): The number of samples that will be propagated through the network at a time. Smaller batch sizes can provide more accurate updates but may require more iterations to complete an epoch.

- Kernel Size (7): The size of the convolutional kernel (filter). A larger kernel size can capture more spatial features but may also increase computational complexity.
- Average Pool Size (3): The size of the window for the average pooling operation. Average pooling reduces the dimensions of the input, helping to decrease the number of parameters and computation necessary.
- Number of Convolutional Layers (2): The number of convolutional layers in the network. More convolutional layers can capture more complex features but may also increase the risk of overfitting and computational cost.
- Loss Function (Categorical cross entropy): The loss function used to optimize the model. Categorical cross entropy is commonly used for classification tasks where the output is a probability distribution between classes.
- Activation Function (Softmax): The activation function used in the output layer. The softmax activation function is typically used for multi-class classification problems as it converts the output logits into probabilities, but is just as useful when working with a binary classification task.
- Optimizer (Adam, Learning Rate = 0.01): The optimizer used to update the model's weights. The Adam optimizer is an adaptive learning rate optimization algorithm that combines the advantages of two other extensions of stochastic gradient descent. The learning rate is set to 0.01.

5.2.3 Overview of Grid Search Results

When looking at the average results of all 576 trained models, the difficulty of the goal we are trying to achieve becomes clear (see table 5.2). Even at the 75% percentile only an accuracy of 55% can be achieved. While this may seem low, it is important to consider the complexity of seizure prediction and the variability in the data. The precision and recall metrics also show a wide range of performance, indicating that some models are better at identifying true positives, while others may have a higher rate of false positives.

The mean ROC AUC score of 0.52 suggests that the models are only slightly better than random guessing on average. However, the maximum ROC AUC score of 0.77 indicates

that there are some configurations that perform significantly better. This highlights the importance of hyperparameter tuning and the potential for further optimization.

Metric	Mean	Std Dev	Min	25%	50%	75%	Max
Accuracy	0.52	0.09	0.23	0.45	0.50	0.55	0.77
Precision	0.45	0.25	0.00	0.36	0.50	0.58	1.00
Recall	0.43	0.33	0.00	0.09	0.41	0.73	1.00
F1 Score	0.40	0.25	0.00	0.15	0.44	0.62	0.81
ROC AUC Score	0.52	0.09	0.23	0.45	0.50	0.55	0.77

Table 5.2: Statistics across all 576 models.

These results underscore the challenges in developing a robust seizure prediction model. The variability in performance metrics suggests that certain hyperparameter configurations are more effective than others. Future work could focus on narrowing down the most promising configurations and exploring additional features or model architectures to improve performance.

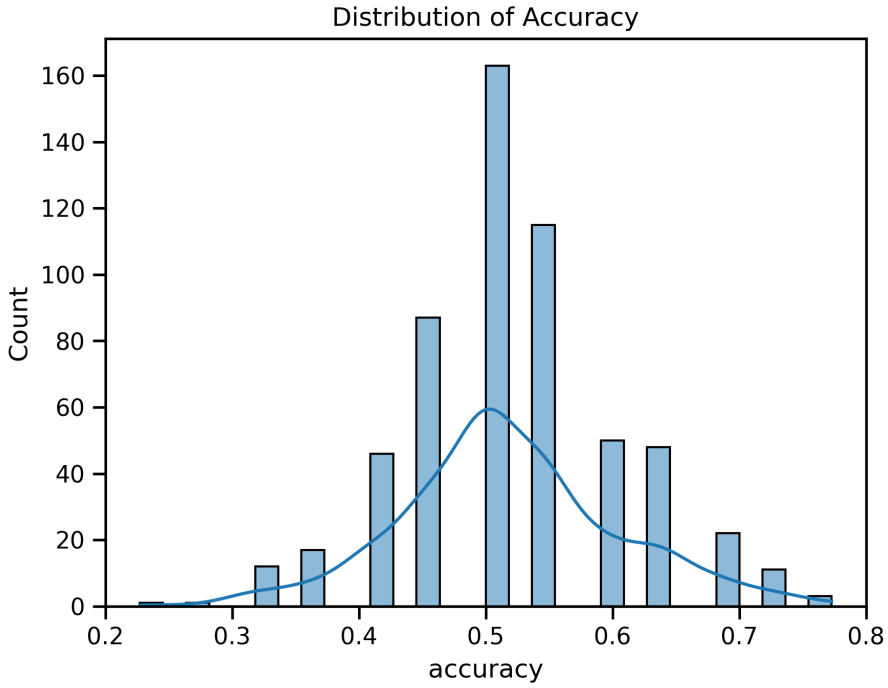


Figure 5.2: Distribution of model accuracy.

The results of the grid search show that the accuracy of the models follows a normal distribution, as depicted in Figure 5.2. This means that most models achieve accuracy close to the average, with fewer models performing significantly better or worse. The bell-shaped curve highlights the consistency in performance across the tested hyperparameter configurations and emphasizes the importance of hyperparameter tuning, as even small adjustments can lead to noticeable improvements in model performance.

5.2.4 Linear Correlation Analysis of Hyperparameters and Performance

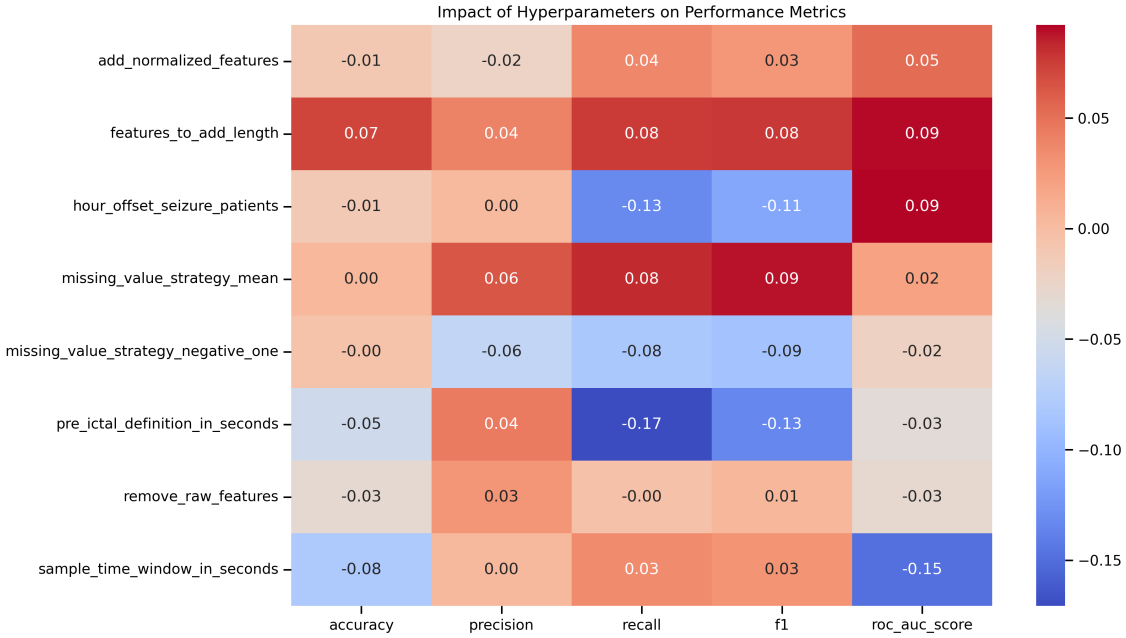


Figure 5.3: Correlation between hyperparameters and performance.

The correlation matrix seen in fig. 5.3 gives us insight into how different hyperparameters influence model performance metrics. When reading the heatmap a positive value, also indicated with colour, shows a correlation to a higher value in the performance metrics when the hyperparameter is increased. In most cases the varied hyperparameters were integers, meaning an increase can be easily understood. In those cases where the hyperparameter was either an array of strings or a boolean value, another approach has to be taken. For example the hyperparameter *features_to_add*, indicated which features should

be added to the dataset. Since an array of strings can hardly be displayed in a correlation heatmap, the length of the array has been used as input for the heatmap. Therefore, we can notice that a longer array of features to be added positively impacts all performance metrics.

For other hyperparameters a single string value was used, for example the missing value strategy. In this case, all possible values of the string have been split into different rows and converted to boolean values. Boolean values can traditionally be displayed as either 0 for false or 1 for true, which again gives us an integer value to work with, which can be displayed in the heatmap. For example, it impacts the performance metrics positively when the missing value strategy of taking the mean is increased. Meaning, the value is increased from 0 to 1 or in other words in switched from false to true. The same principle can be applied to all hyperparameters that have been converted to boolean values for the sake of visualizing the correlation.

Overall, most hyperparameters show only slight correlations with these metrics, indicating that their effects might be more complex and dependent on interactions rather than direct linear relationships. However, some patterns stand out.

One notable observation is that the pre-ictal definition in seconds has a small but noticeable negative correlation with recall (-0.17) and F1-score (-0.13). This suggests that as the pre-ictal period increases, the ability of the model to correctly identify seizures decreases. This could be due to longer pre-ictal periods introducing more noise, making it harder for the model to distinguish between pre-ictal and non-pre-ictal states. Since recall is directly tied to F1-score, we see a similar drop in F1 as well.

Similarly, sample time window in seconds shows a small negative correlation with ROC AUC score (-0.15). This might indicate that larger time windows introduce variability that reduces the models' ability to differentiate between classes effectively. If the time window is too broad, the model might struggle to detect meaningful patterns that contribute to classification performance. This goes against what has been shown in previous research using EEG data, where a 30 to 60-minute time window has been proven optimal.

Interestingly, the inclusion of additional normalized features and removal of raw features appears to have almost no significant correlation with performance. This suggests that normalizing features or removing raw data does not drastically change the models' behaviour linearly, though it might still have nonlinear effects.

We can make another interesting observation from the missing value strategy. The strategy of taking the mean of a feature in case of missing values shows a slight advantage, leading to an improvement of 0.09 in the F1 score compared to imputing negative one. This indicates that mean imputation might be a more effective approach for handling missing values in this context, though its impact could still depend on specific datasets and interactions with other hyperparameters.

Overall, while no hyperparameter shows a strong, direct relationship with performance, the small negative correlations with pre-ictal definition and sample time window suggest that optimizing these parameters carefully could lead to marginal improvements in recall and AUC. The weak correlations across the board also highlight that performance likely depends on complex interactions rather than single hyperparameters alone.

Metric	**Top 10%**	**Bottom 10%**
Accuracy	0.68	0.37
Precision	0.68	0.25
Recall	0.77	0.27
F1 Score	0.69	0.25
ROC AUC Score	0.67	0.37
Add Normalized Features	0.56	0.56
Features to Add Length	3.79	3.58
Hour Offset Seizure Patients	1.84	3.16
Pre-Ictal Definition (seconds)	570.53	702.11
Remove Raw Features	0.37	0.44
Sample Time Window (seconds)	642.11	810.53
Missing Value Strategy (Mean)	0.51	0.51
Missing Value Strategy (Negative one)	0.49	0.49

Table 5.3: Hyperparameter trends for high and low performing models.

The table in table 5.3 highlights the trends in hyperparameters for the top 10% and bottom 10% performing models. Interestingly, the inclusion of normalized features does not show a significant difference between high and low performing models, suggesting that this factor alone does not drastically impact performance.

The length of features to add is slightly higher in top-performing models, indicating that a richer feature set might contribute to better performance. The hour offset for seizure patients is lower in top-performing models, suggesting that a smaller offset might be

beneficial. Similarly, the pre-ictal definition and sample time window are shorter for top-performing models, which aligns with the earlier observation that longer periods might introduce noise and reduce model effectiveness.

The strategies for handling missing values do not show a significant difference between high and low performing models, indicating that the choice between mean imputation and using a negative one placeholder does not have a strong linear impact on performance. Overall, these trends provide valuable insights into which hyperparameters might be fine-tuned to achieve better model performance.

5.2.5 Identifying the best model

Having completed training and testing of one model type leaves us with metrics and results for 576 differently configured models. The best performing model does not necessarily need to be the best in every category. While accuracy and precision may be high for some model, it may lack in the recall metric. For this use case we deem the best model to be the one with the highest accuracy since we value overall correctness of the model. While in this context one could argue that it is more important to not miss any seizures, a first look at the results showed that some models simply classify every sample as pre-ictal and therefore have a recall value of 1. Since we do not want to deem a model like this as the best, we use accuracy as the defining metric. All other mentioned metrics will also be included when presenting the results.

We did not train a baseline model to compare the results of the best performing model to a dummy provide a sufficient overview of the average performance of the models. The detailed results of the best performing model are presented in the following section.

5.2.6 Performance of the best model

The CNN model with the highest accuracy demonstrates a balanced performance across various metrics. The accuracy of the model is 0.7727, indicating that it correctly predicts the outcome approximately 77% of the time. The precision is 0.7143, suggesting that the model, while good at identifying positive cases, might too eagerly classify positive samples as such. The confusion matrix of the model will give more insight later on. The

recall is 0.9091, indicating that the model is very performant at correctly predicting positive cases. The F1 Score, showing the harmonic mean of precision and recall, is 0.8, reflecting the models consistent performance. These metrics indicate that the CNN model provides a well-rounded performance with a strong recall capability.

Metric	Value
Accuracy	0.7727
Precision	0.7143
Recall	0.9091
F1 Score	0.8

Table 5.4: Performance metrics for the best performing CNN model.

Parameter	Value	Parameter	Value
Marker Type	All	Pre-Ictal Definition	2 min
Seizure Types	All	Post-Ictal Definition	10 min
Overlapping Markers Allowed	No	Sampling Rate	0.5 Hz
Exclude Seizure Start Markers	No	Faulty Data Value Cut-off	50,000
Exclude Seizure End Markers	Yes	Add Normalized Features	No
No-Seizure Data Percentage	0%	Remove Raw Features	Yes
Hour Offset For Inter-Ictal Markers	5 hours	Additional Features	SD, Var, Mean, Min, Max
Pre-Ictal Label Percentage	50%	Missing Value Strategy	Negative-One
Inter-Ictal Label Percentage	50%	Add Time of Day Feature	Yes
Ictal Label Percentage	0%	Remove Top Percentile (Outliers)	2%
Post-Ictal Label Percentage	0%	Train-Test Strategy	Last Seizures
Sample Time Window	5 min		

Table 5.5: Parameters for best performing CNN model.

The best performing model uses a sample time window of 5 minutes, which is in line with the trend seen in the average, where the top performing models use shorter time windows. The hour offset for inter-ictal markers has been chosen as 5 hours for the best performing model, this is also similar to the average which showed that a lower offset leads to better results. One has to keep in mind that the average for the offset is skewed a bit, due to the fact that we included a 10-hour offset in the grid-search. It also uses the full range of features for its prediction, including the mean, minimum and maximum for each feature. For the missing value strategy negative-one has proven to be the best, even though the average showed that this hyperparameter did not have a great impact on the performance of the models. No normalized features were added, and raw features were also removed. This means that only the additional features (standard deviation, variance, mean, minimum and maximum) were used for prediction.

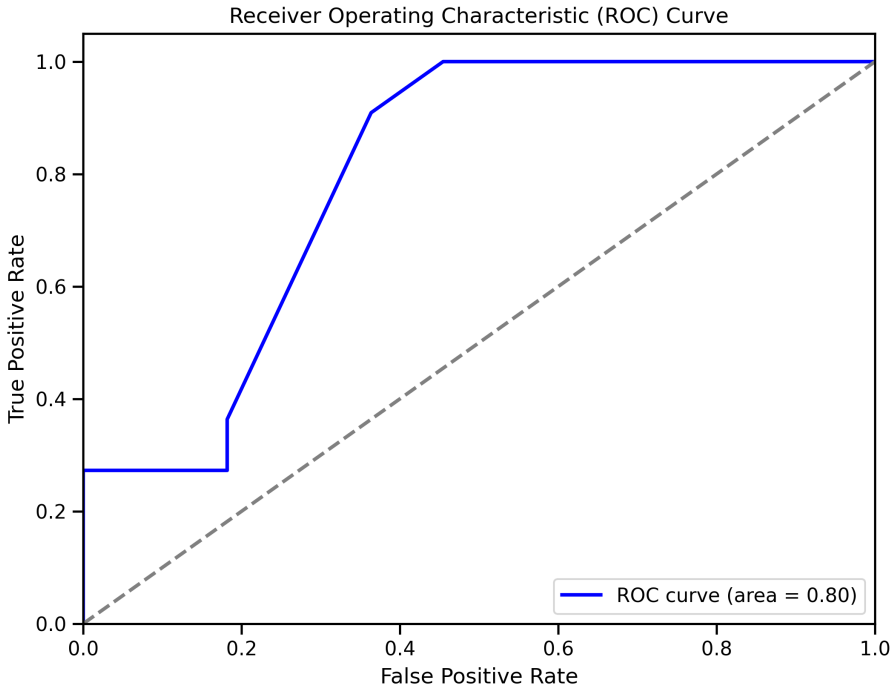


Figure 5.4: ROC Curve for the best performing CNN model.

The ROC curve shown in fig. 5.4 suggests that the model is performing well, with an AUC of 0.80, meaning it effectively distinguishes between positive and negative cases about 80% of the time. While this is a strong result, robustness tests need to be performed to accurately assess whether this performance would be replicable on unseen data.

The confusion matrix shows how well the model distinguishes between pre-ictal and non-pre-ictal states. It correctly identifies 10 pre-ictal cases while misclassifying 1, giving it a recall of 91%, meaning it rarely misses a true pre-ictal case.

However, it misclassifies 4 non-pre-ictal cases as pre-ictal, leading to a specificity of 64%. This indicates a tendency for false positives. While this is still problematic, in the case of epilepsy a false alarm are not as costly as not identifying a seizure correctly. Precision is 71%, meaning 3 out of 10 predicted pre-ictal cases are incorrect. The model prioritizes sensitivity over specificity, making it effective for detecting pre-ictal cases but at the cost of some false alarms. In medical applications, this trade-off may be acceptable, but reducing false positives could improve usability. Adjusting the classification threshold or processing features differently may help achieve better balance.

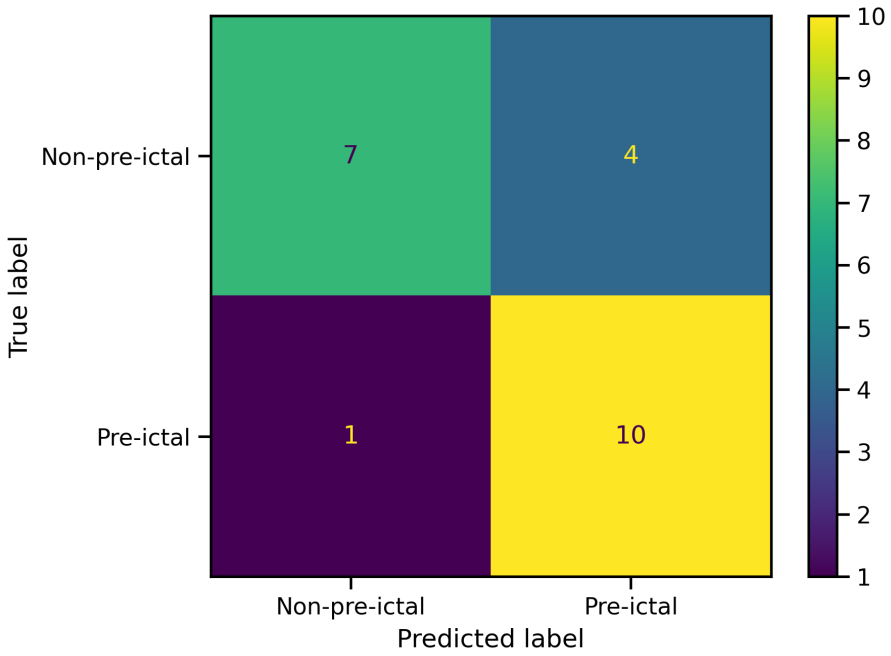


Figure 5.5: Confusion Matrix for the best performing CNN model.

The training accuracy chart shows a problem, the accuracy goes down again after having reached 70% after epoch 200. This is due to overfitting, where the model starts to memorize the training data instead of learning to generalize from it. To mitigate this, techniques such as early stopping, regularization, or dropout could be employed to improve the generalization capability of the model. Additionally, the loss levels out at a very high rate of about 0.45, indicating that the model is not learning effectively after a certain point. This plateau suggests that the model has reached its capacity to learn from the training data and is not improving further. To address this, further tests could experiment with different learning rates, batch sizes, or even try different architectures to see if they provide better convergence and lower loss values.

5.2.7 Predictions of the best model

As we have the meta-data for all seizures that are fed into the model we can take a look which types of seizures are being misidentified by the model. The best model, according to accuracy, only missed one pre-ictal sample, which had a non-motor onset. While it did

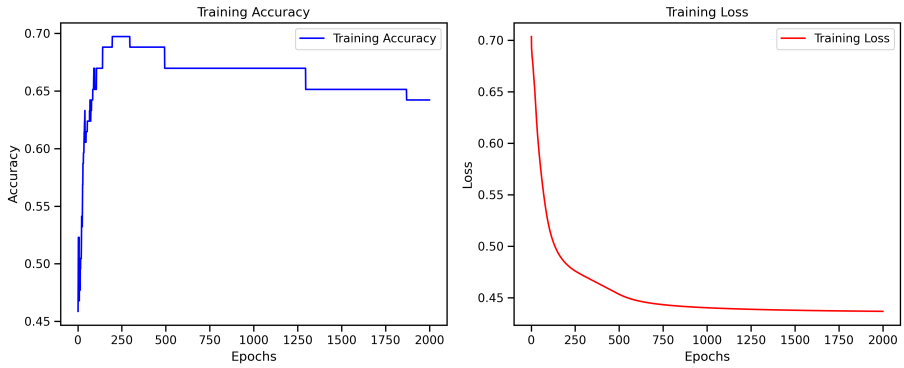


Figure 5.6: Training loss and accuracy for the CNN model.

classify other non-motor onset seizure correctly, this specific seizure type categorization seen in table 5.6 is only found once in the test set.

Table 5.6: Prediction accuracy on ictal test cases by seizure type

Seizure Type (Epilepsia IDs)	**Correct**	**Total**	**Accuracy (%)**
Focal Onset - Aware - Motor Onset - tonic (4SCV79, CEU856, HAES28)	7	7	100.0
Focal Onset - Impaired Awareness - Nonmotor Onset - autonomic (DJDEH3, LFLWD7)	4	4	100.0
Focal Onset - Impaired Awareness - Motor Onset - automatisms (H6X6G9)	1	1	100.0
Focal Onset - Aware - Motor Onset - automatisms (HAES28)	1	1	100.0
Focal Onset - Aware - Nonmotor Onset - autonomic (XY5FE5)	0	1	0.0
Total	13	14	92.9

For the inter-ictal test samples the model reacted sensitively, only achieving 63% accuracy when identifying negative samples. Since we are talking about negative inter-ictal samples here, we could only match the Epilepsia ID to the seizure type the patient usually experiences. However, in most cases, a singular patient within our study experienced

multiple types of seizures if they experienced more than one at all. For this reason, only the Epilepsia ID is listed in the first column, without any association to a specific seizure type.

Table 5.7: Prediction accuracy on inter-ictal test cases

Epilepsia ID	Correct	Total	Accuracy (%)
CEU856	1	1	100.0
H6X6G9	0	1	0.0
ZA8OX5	1	1	100.0
LFLWD7	1	3	33.3
DJDEH3	1	2	50.0
C4TQG5	1	1	100.0
4SCV79	1	1	100.0
9E4831	1	1	100.0
Total	7	11	63.6

5.2.8 Robustness of the best model

While these results seem promising, the model might simply be overfitting or reacting to random noise. To rule out these scenarios, many tests regarding changing the input data and looking at the reaction of the model can be conducted. We will limit ourselves to label shuffling, feature permutation, feature perturbation and feature removal tests.

Label Shuffling

The label shuffling test changes the labels of the test dataset but keeps the input data the same when evaluating the model. If the model performs significantly worse on the permuted labels compared to the original ones, it suggests that the model is capturing meaningful patterns rather than overfitting to noise. Naturally, this test was conducted with retraining of the model.

As one can see in fig. 5.7, changing the labels leads to completely non-functional model that does not correctly identify pre-ictal states whatsoever. This leads to a significant drop

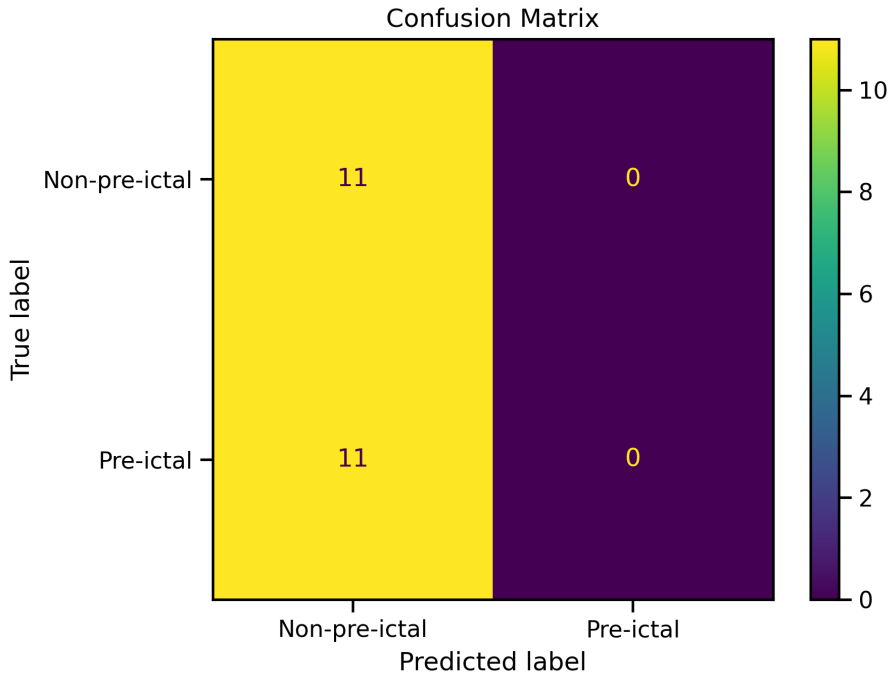


Figure 5.7: Confusion matrix for shuffled label test for the best performing CNN model.

in performance metrics as shown in section 5.2.8. The accuracy drops to 0.5, which is equivalent to random guessing. Precision, recall, and F1 score all drop to 0.0, indicating that the model fails to identify any positive cases correctly. The ROC AUC score also drops to 0.3182, further confirming that the model is not performing better than random chance. While these results suggest that the original models' performance is not due to overfitting or noise, more robustness tests need to be conducted to fully confirm the validity of the model.

Metric	Value
Accuracy	0.5
Precision	0.0
Recall	0.0
F1 Score	0.0
ROC AUC Score	0.3182

Table 5.8: Performance metrics for best performing CNN model with shuffled labels.

Feature Permutation

To further assess the robustness of the CNN model, feature permutation tests were conducted. These tests involve shuffling each feature one after another and trying to predict on the modified test dataset. The drop in accuracy then signifies the impact the feature has on the performance of the model. The higher the value the higher the drop in accuracy. These tests happen without retraining the model, only the input test data gets altered. Figure 5.8 shows the calculated importances for the most impactful features with a value above 0.05.

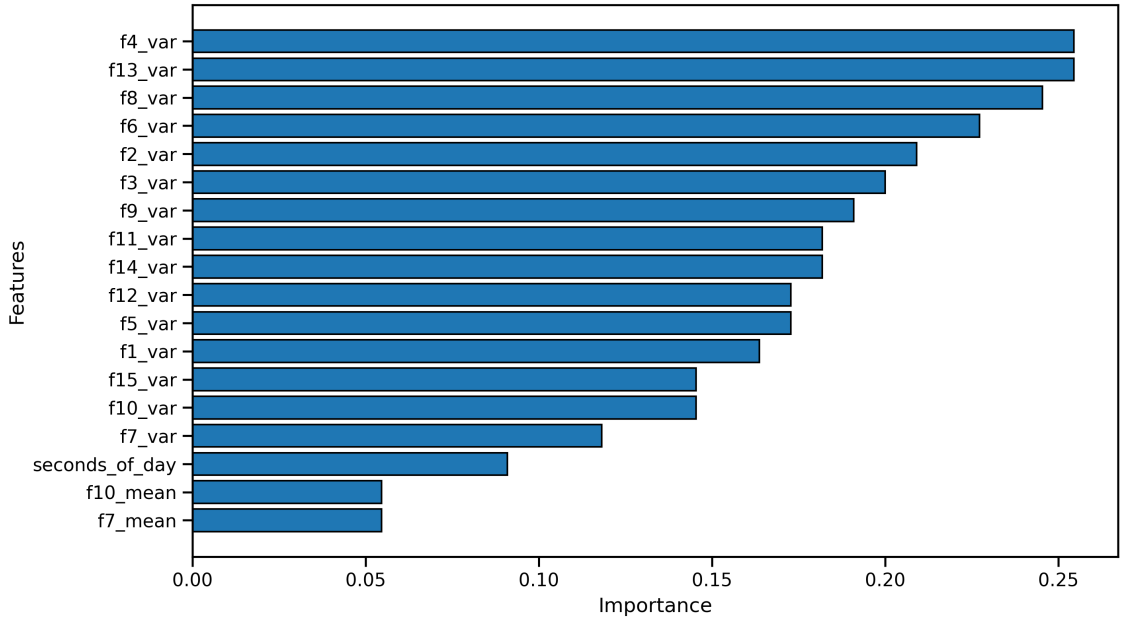


Figure 5.8: Feature importances for the best performing CNN model. Importance score shows the average accuracy drop when the feature gets randomly shuffled. Importance scores below 0.05 not shown.

As can be seen in fig. 5.8, variance features have a great impact on the performance of the model. Since no normalized features were added in this model and on top of that raw features were removed, it could be reasoned that variance naturally should have a high impact on performance. However, then the question arises why only variance plays a significant role in this model and not also the standard deviation and possibly other metrics like the mean, minimum and maximum of each feature. All variance features, 1

through 15 can be seen as strongly impactful, with not a single different feature coming close. Additionally, the meaning of the variation between variance features importance scores could be questioned, i.e. what should be we take away from the fact the f4 feature has an importance score double as high as f7. The role of the *seconds_of_day* feature having an importance score of just less than 0.1 needs to be investigated further. The results of this permutation importance analysis builds upon what we have seen in the linear correlation hyperparameter analysis done across all trained models.

Feature Perturbation

Feature perturbation tests are designed to evaluate the robustness of a model by introducing controlled noise or alterations to specific features in the input data. The goal is to observe how the models predictions change in response to these perturbations, which can provide insights into the models reliance on particular features and its ability to generalize.

For this test we try to disturb the dataset using three approaches. In the first we increase a single feature by 20% and then try to predict on the modified input data. We repeat this for each feature and assess the drop in accuracy. For the second approach we do the same, but this time decreasing each feature by 20%. With the third approach we add normally distributed Gaussian noise to the input data to see how the model reacts to different levels of noise. The Gaussian noise is proportional to the standard deviation of the whole feature across all samples.

Figure 5.9 shows the impact increasing features by 20% had on the accuracy of the model. The sensitivity range displays a drop in accuracy in the positive range and an increase in accuracy in the negative range. Only the features where the increase made a difference in the resulting accuracy are shown.

Interestingly, only 5 features lead to a different accuracy score than the original input dataset. Of those 5, only the increase for the *seconds_of_day* feature resulted in a performance drop of 0.05. In all other cases, two mean features, one variance and one minimum, the increase resulted in a 0.05 better accuracy than the unmodified dataset.

Looking at the inverse test, seen in fig. 5.10 we have 7 features that resulted in either a drop or increase in performance. Decreasing 6 of these features lead to a performance

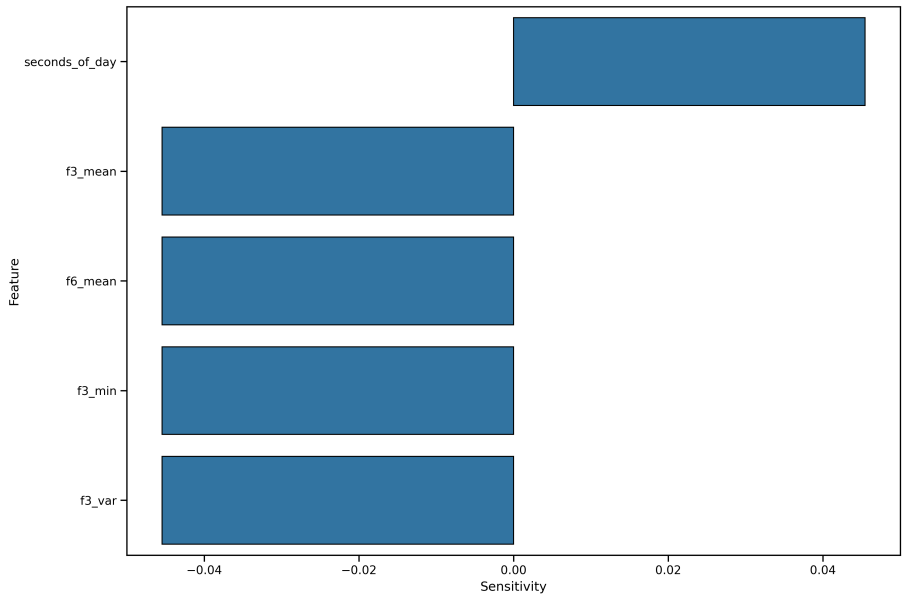


Figure 5.9: Performance impact when increasing features by 20%. Positive sensitivity values indicate a drop in accuracy, negative sensitivity values an increase in accuracy.

drop of 0.05, measured in accuracy. Only decreasing the variance for f4 resulted in a 0.05 increase in accuracy.

The results of model when Gaussian noise (see fig. 5.11) is introduced shows the instability of the model. Even before reaching a noise level of 0.1 the model becomes only slightly better than random guessing, which indicates possible overfitting. The sudden increase between the noise level of 0.3 and 0.4 should be attributed to spurious correlation.

Feature Removal

With this test we want to check which features are most important for the model in the training and therefore which features are being learnt from the most. Additionally, we want to check how robust the model is to removing features before training. If the model breaks down when removing any feature it might indicate overfitting.

The feature ablation test involves removing one feature at a time from the training and testing dataset and observing the impact on the model's performance. The result of this test is visualized in fig. 5.12. Some features can be identified as not having a strong impact

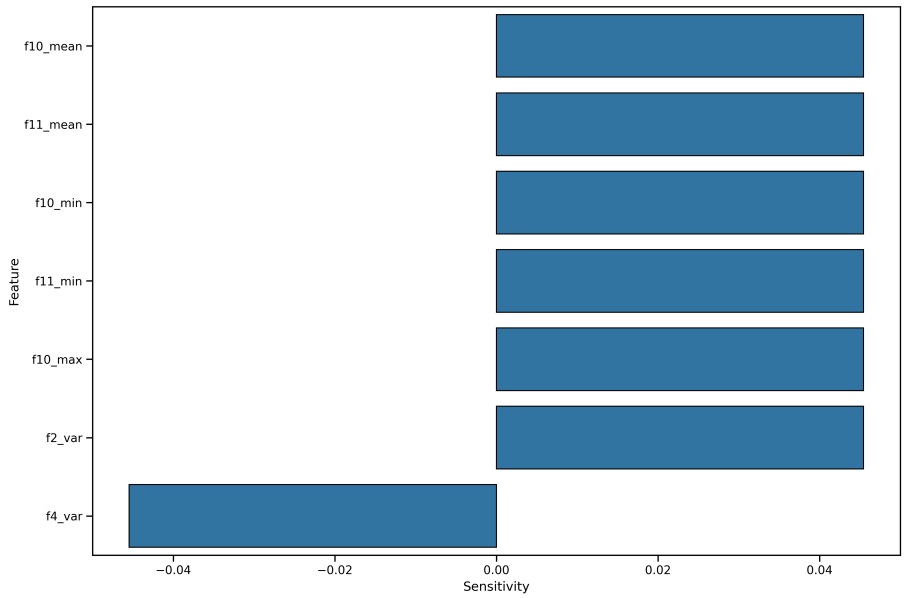


Figure 5.10: Performance impact when decreasing features by 20%. Positive sensitivity values indicate a drop in accuracy, negative sensitivity values an increase in accuracy.

on learning, such as the minimum and maximum features of f14 and f8 respectively. However, in most cases the model takes a hit in accuracy when any singular feature is removed. For over half of all features the removal leads to a performance equivalent to random guessing or worse. It is not plausible to say that every feature is of high importance to the model, which is why the performance breaks down. The more likely reason is that the model is overfitting on either the training data or the hyperparameter configuration.

One more compelling reason why the model's performance might be spurious correlation, is the loss in performance when the *seconds_of_day* feature is removed. It is unlikely that the feature should have such a high impact on the model. As was evaluated in section 4.3.1, there was no indication that the seizure times correlated with the time of day.

When looking at the averages of the new resulting accuracies when certain feature types are dropped, it can be seen that in most cases the model loses any classification ability with its accuracy dropping to about 50%.

These results, especially including the Gaussian noise tests, suggest that the model is overfitting to the hyperparameters because it relies heavily on specific configurations of

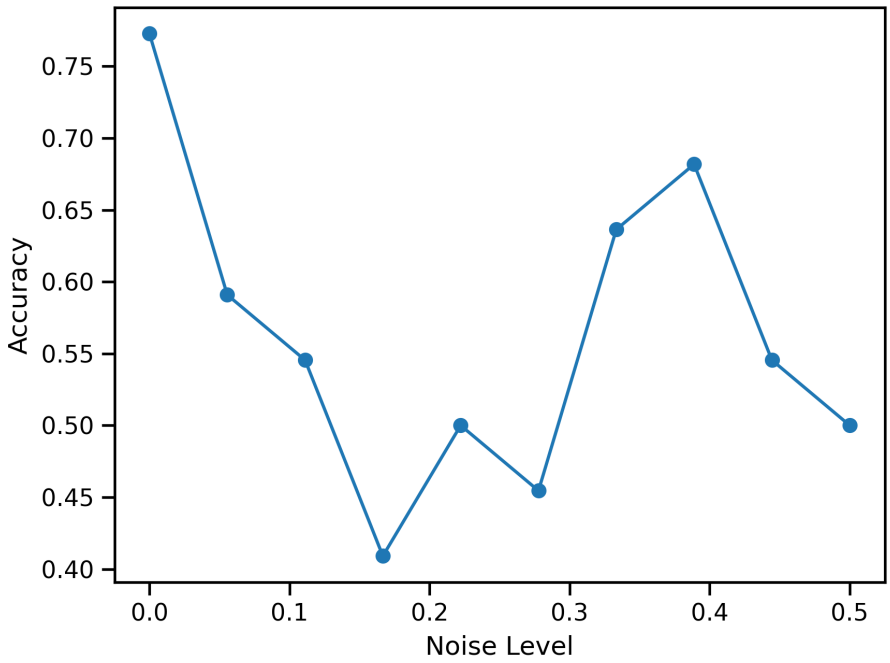


Figure 5.11: Impact of Gaussian noise on model performance. Noise level calculated as fraction of standard deviation.

the input data and feature set to achieve high performance. This dependency indicates that the model is not generalizing well to unseen data and is instead learning patterns that are specific to the training set. While overfitting can occur due to overly complex model architectures, we do not believe this to be the case. Insufficient training data is the most likely reason why this is happening. Lack of regularization techniques such as dropout could additionally be another reason for this drop in performance. To address this more experiments with a larger dataset or additional regularization methods could help improve the models robustness and generalization capability.

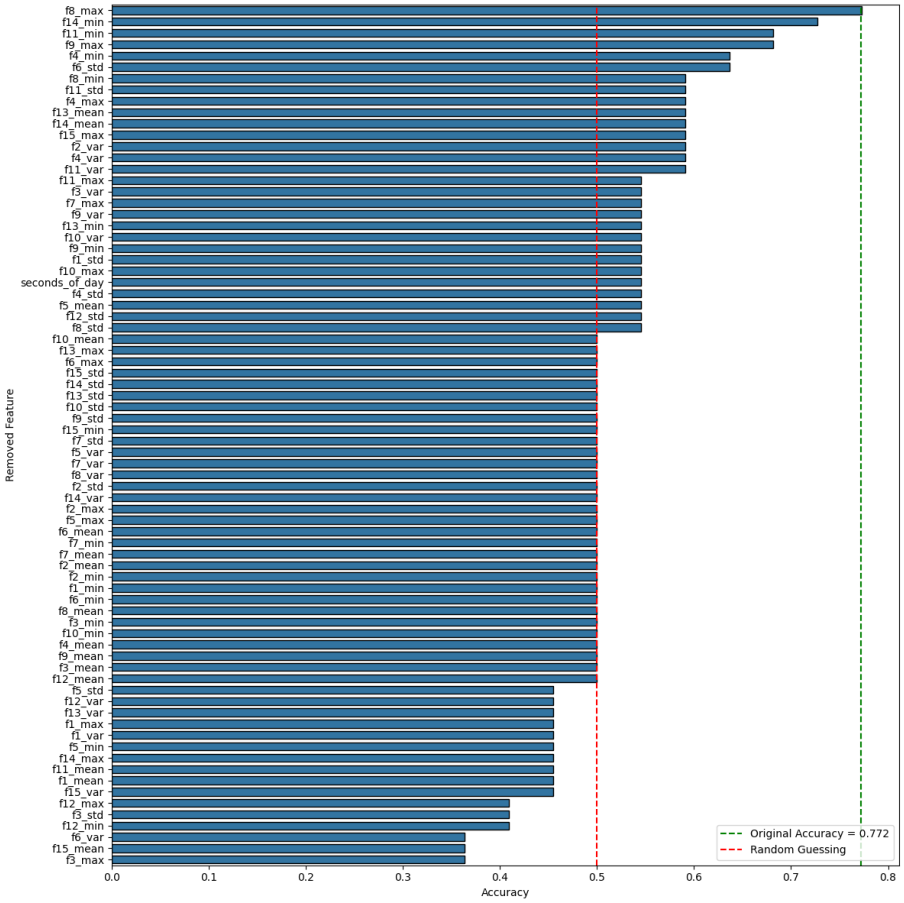


Figure 5.12: Impact of feature removal on CNN model performance. Accuracy is shown as the new resulting performance on the test set. The green line marks the performance of the original model, the red line marks where performance drops to random guessing.

Removed Feature Type	Resulting Accuracy (%)
Overall Average	51.9
Variance Features	50.6
Standard Deviation Features	51.8
Seconds of Day Feature	54.5
Mean Features	50.0
Maximum Features	53.0
Minimum Features	53.9

Table 5.9: Average accuracy for the different feature types from results seen in fig. 5.12.

6 Conclusion

6.1 Limitations

There are several aspects that could be limiting potential effectivity of the resulting classification models. The final dataset of valuable patients and seizures is small compared to other medical research areas where the number of total samples used in a machine learning process are often in the thousands [79]. This probably does affect the resulting performance of all machine learning models that have been trained using our data. It cannot be determined how much of an improvement could be made using more VOC data.

On top of the possibility that not enough data has been captured, VOC data can also be inherently noisy, influenced by a range of factors such as environmental contamination, sensor malfunctions, and other external variables. These contaminants might come from background sources like food, cleaning products, or even air quality variations [80, 81]. Such external noise can obscure meaningful patterns tied to seizure events, making it harder to differentiate between seizure-related VOC changes and those arising from unrelated sources. Furthermore, VOC sensors themselves can introduce noise due to imperfections in calibration or drift over time, which can impact the accuracy and consistency of data over extended periods.

For supervised learning models, accurately labelled data is essential. In the case of seizure prediction, the only reference point is a seizure marker, which could be faulty or misleading. Additionally, VOC profiles may not change in a consistent or easily detectable way before or during a seizure. Inaccurate or imprecise labelling of data can impair model performance, as the model learns from potentially incorrect or incomplete information.

The VOC profiles associated with seizures can vary significantly across individuals due to differences in metabolic processes, medications, genetics, or even lifestyle factors like diet.

This makes it difficult to generalize findings from one patient or population to another. If a model is trained on VOC data from a single individual or a small group, it may not perform as well when applied to a broader or different population. The variability in VOC signatures can also complicate the identification of consistent patterns associated with seizures [82].

Currently, there are no standardized protocols for collecting or interpreting VOC data in the context of seizure prediction. Differences in sensor types, environmental conditions, and data preprocessing steps can lead to significant inconsistencies between datasets and experiments. Without standardization, it becomes difficult to compare results across different studies or to combine data from multiple sources. This can hinder progress in the field and limit the ability to draw generalizable conclusions.

The models, especially the well performing models, are not above questioning, while the label shuffling test showed that the model is learning actual patterns found in the data, the robustness tests highlighted clear weaknesses when features were dropped completely. The introduction of only little noise to the dataset also lead to the model losing a significant portion of its accuracy. It remains unclear whether the model overfitted on the hyperparameter and just by chance performed well on the test data. More statistical tests or if possible an actual real-world test in the hospital would be the best option to provide evidence that the model is predicting certain seizures correctly.

The selection of the parameter grid inherently limits the study to some degree. While all decisions regarding the parameter grid were based on state-of-the-art literature and in some cases preliminary experiments, it is possible that well performing combinations might have been missed. Restraints regarding time and computing power limited the number of possible combinations that we were able to try. Further parameter range tests could possibly lead to interesting results, as we already saw positive correlation trends regarding certain hyperparameter.

6.2 Outlook

It remains unclear whether a prediction based only on volatile organic compound data can be made. While the grid-search resulted in a few well performing models, with the best model reaching an accuracy of 77%, those models did not fully withstand thorough

robustness tests when changing or dropping out random features. While we cannot say for certain how those trained models would perform in a clinical setting predicting unseen seizures, we do not expect them to achieve similar performance as on the testing data. The average across all trained models showed that the task at hand remains a difficult one, seeing as only the top quartile achieve reasonably good results.

This is not to say that it is not viable to predict epileptic seizures using olfactory data, but that in this case the combination of limitations we experienced during the study resulted in unstable models that in theory perform well, but might be overfitting on the combination of hyperparameter or on the training data itself.

Regarding the importance of hyperparameter, we did see small trends indicating some value ranges were better than others. While these trends were not strong enough to draw definitive conclusions, they do suggest that further investigation into hyperparameter optimization could be beneficial. Future work should focus on more robust methods for feature selection and model validation to ensure that the models are not overfitting and can generalize well to unseen data.

In the feature importance analysis, the variance of all features emerged as a significant factor, whereas other metrics such as standard deviation, mean, minimum, and maximum had little impact on the model's performance. Feature ablation tests, where the model was retrained after removing individual features, revealed that while a small number of features had minimal effect on performance, in most cases, removing a feature resulted in a significant drop in accuracy. This suggests that the initial accuracy may have been influenced by spurious correlations.

In conclusion, while our study faced several challenges and limitations, it has provided valuable insights into the potential and limitations of using volatile organic compound data for predicting epileptic seizures. Not only was this a first try at VOC-based prediction of seizures but also a first investigation into which hyperparameters have a significant impact on the results, something that has not been studied thoroughly yet. Further research with larger datasets, more robust validation techniques, and improved feature selection methods is necessary to fully understand and harness the predictive power of this approach.

7 Appendix

Please note that the repository containing all code and results is not publicly due to the sensitive nature of the data involved. If you are interested in the code please contact me at me@alexanderwolf.xyz.

Bibliography

[1] Shekhar Saxena and Shichuo Li. "Defeating epilepsy: A global public health commitment". In: *Epilepsia Open* 2.2 (Mar. 2017), pp. 153–155. ISSN: 2470-9239. DOI: 10.1002/epi4.12010. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5719859/ (visited on 02/22/2024) (cit. on p. 1).

[2] World Health Organization. "Epilepsy: a public health imperative". ar. In: (2019). URL: https://iris.who.int/handle/10665/325440 (visited on 03/09/2024) (cit. on p. 1).

[3] Marzieh Savadkoohi, Timothy Oladunni, and Lara Thompson. "A machine learning approach to epileptic seizure prediction using Electroencephalogram (EEG) Signal". In: *Biocybernetics and Biomedical Engineering* 40.3 (July 2020), pp. 1328–1341. ISSN: 0208-5216. DOI: 10.1016/j.bbe.2020.07.004. URL: https://www.sciencedirect.com/science/article/pii/S0208521620300851 (visited on 11/10/2024) (cit. on p. 1).

[4] Buajieerguli Maimaiti et al. "An Overview of EEG-based Machine Learning Methods in Seizure Prediction and Opportunities for Neurologists in this Field". In: *Neuroscience* 481 (Jan. 2022), pp. 197–218. ISSN: 0306-4522. DOI: 10.1016/j.neuroscience.2021.11.017. URL: https://www.sciencedirect.com/science/article/pii/S0306452221005765 (visited on 11/10/2024) (cit. on p. 1).

[5] Brian Litt and Javier Echauz. "Prediction of epileptic seizures". In: *The Lancet Neurology* 1.1 (May 2002), pp. 22–30. ISSN: 1474-4422. DOI: 10.1016/S1474-4422(02)00003-0. URL: https://www.sciencedirect.com/science/article/pii/S1474442202000030 (visited on 01/08/2025) (cit. on p. 2).

[6] William C. Stacey. "Seizure Prediction Is Possible-Now Let's Make It Practical". eng. In: *EBioMedicine* 27 (Jan. 2018), pp. 3–4. ISSN: 2352-3964. DOI: 10.1016/j.ebiom.2018.01.006 (cit. on p. 2).

[7] Arthur W. Barrios, Pablo Sánchez-Quinteiro, and Ignacio Salazar. "Dog and mouse: toward a balanced view of the mammalian olfactory system". In: *Frontiers in Neuroanatomy* 8 (Sept. 2014), p. 106. ISSN: 1662-5129. DOI: 10.3389/fnana.2014.00106. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4174761/ (visited on 01/10/2025) (cit. on p. 2).

[8] Bogusław Buszewski et al. "Identification of volatile lung cancer markers by gas chromatography–mass spectrometry: comparison with discrimination by canines". en. In: *Analytical and Bioanalytical Chemistry* 404.1 (July 2012), pp. 141–146. ISSN: 1618-2650. DOI: 10.1007/s00216-012-6102-8. URL: https://doi.org/10.1007/s00216-012-6102-8 (visited on 01/10/2025) (cit. on p. 2).

[9] Paula Jendrny et al. "Canine olfactory detection and its relevance to medical detection". In: *BMC Infectious Diseases* 21.1 (Aug. 2021), p. 838. ISSN: 1471-2334. DOI: 10.1186/s12879-021-06523-8. URL: https://doi.org/10.1186/s12879-021-06523-8 (visited on 01/09/2025) (cit. on p. 2).

[10] Edward Maa et al. "Canine detection of volatile organic compounds unique to human epileptic seizure". In: *Epilepsy & Behavior* 115 (Feb. 2021), p. 107690. ISSN: 1525-5050. DOI: 10.1016/j.yebeh.2020.107690. URL: https://www.sciencedirect.com/science/article/pii/S1525505020308702 (visited on 10/16/2024) (cit. on pp. 2, 16, 17, 28, 32).

[11] E. H. Maa, J. Arnold, and C. K. Bush. "Epilepsy and the smell of fear". In: *Epilepsy & Behavior* 121 (Aug. 2021), p. 108078. ISSN: 1525-5050. DOI: 10.1016/j.yebeh.2021.108078. URL: https://www.sciencedirect.com/science/article/pii/S1525505021003127 (visited on 10/16/2024) (cit. on pp. 2, 17).

[12] Grace C. Luff et al. "The role of trained and untrained dogs in the detection and warning of seizures". In: *Epilepsy & Behavior* 150 (Jan. 2024), p. 109563. ISSN: 1525-5050. DOI: 10.1016/j.yebeh.2023.109563. URL: https://www.sciencedirect.com/science/article/pii/S1525505023004821 (visited on 10/16/2024) (cit. on pp. 2, 16).

[13] Amélie Catala et al. "Dogs demonstrate the existence of an epileptic seizure odour in humans". In: *Scientific Reports* 9 (Mar. 2019), p. 4103. ISSN: 2045-2322. DOI: 10.1038/s41598-019-40721-4. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6438971/ (visited on 10/16/2024) (cit. on pp. 2, 16).

[14] Bendong Zhao et al. “Convolutional neural networks for time series classification”. In: *Journal of Systems Engineering and Electronics* 28.1 (Feb. 2017). Conference Name: Journal of Systems Engineering and Electronics, pp. 162–169. ISSN: 1004-4132. DOI: 10.21629/JSEE.2017.01.18. URL: https://ieeexplore.ieee.org/document/7870510 (visited on 03/12/2025) (cit. on pp. 3, 65).

[15] Sandra Isaza-Jaramillo et al. “The abbreviation “PWE” may carry a negative connotation compared with the labels “person with epilepsy” and “epileptic””. In: *Seizure* 76 (Mar. 2020), pp. 167–172. ISSN: 1059-1311. DOI: 10.1016/j.seizure.2020.02.010. URL: https://www.sciencedirect.com/science/article/pii/S1059131120300510 (visited on 11/05/2024) (cit. on p. 5).

[16] Paula T. Fernandes, Nelson F. de Barros, and Li M. Li. “Stop saying epileptic”. eng. In: *Epilepsia* 50.5 (May 2009), pp. 1280–1283. ISSN: 1528-1167. DOI: 10.1111/j.1528-1167.2008.01899.x (cit. on p. 5).

[17] Markus Reuber. “"Epileptics", "people with epilepsy", "PWE", "epilepsy patients"–what is the best label?” eng. In: *Seizure* 23.5 (May 2014), p. 327. ISSN: 1532-2688. DOI: 10.1016/j.seizure.2014.01.014 (cit. on p. 5).

[18] GBD 2016 Epilepsy Collaborators. “Global, regional, and national burden of epilepsy, 1990-2016: a systematic analysis for the Global Burden of Disease Study 2016”. eng. In: *The Lancet. Neurology* 18.4 (Apr. 2019), pp. 357–375. ISSN: 1474-4465. DOI: 10.1016/S1474-4422(18)30454-X (cit. on p. 6).

[19] Mikhail V. Blagosklonny. “No limit to maximal lifespan in humans: how to beat a 122-year-old record”. In: *Oncoscience* 8 (Dec. 2021), pp. 110–119. ISSN: 2331-4737. DOI: 10.18632/oncoscience.547. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8636159/ (visited on 03/07/2024) (cit. on p. 6).

[20] Jessica Falco-Walter. “Epilepsy—Definition, Classification, Pathophysiology, and Epidemiology”. en. In: *Seminars in Neurology* 40.06 (Dec. 2020), pp. 617–623. ISSN: 0271-8235, 1098-9021. DOI: 10.1055/s-0040-1718719. URL: http://www.thieme-connect.de/DOI/DOI?10.1055/s-0040-1718719 (visited on 03/06/2024) (cit. on p. 7).

[21] Anthony K. Ngugi et al. “Estimation of the burden of active and life-time epilepsy: a meta-analytic approach”. eng. In: *Epilepsia* 51.5 (May 2010), pp. 883–890. ISSN: 1528-1167. DOI: 10.1111/j.1528-1167.2009.02481.x (cit. on p. 7).

[22] World Health Organization. "Epilepsy: a public health imperative". ar. In: (2019). Accepted: 2019-06-20T09:53:27Z Number: WHO/MSD/MER/19.2 Publisher: World Health Organization. URL: https://iris.who.int/handle/10665/325440 (visited on 03/09/2024) (cit. on p. 7).

[23] Devender Bhalla et al. "Etiologies of epilepsy: a comprehensive review". en. In: *Expert Review of Neurotherapeutics* 11.6 (June 2011), pp. 861–876. ISSN: 1473-7175, 1744-8360. DOI: 10.1586/ern.11.51. URL: http://www.tandfonline.com/doi/full/10.1586/ern.11.51 (visited on 03/20/2025) (cit. on p. 7).

[24] Magdalena Bosak et al. "Implementation of the new ILAE classification of epilepsies into clinical practice — A cohort study". English. In: *Epilepsy & Behavior* 96 (July 2019). Publisher: Elsevier, pp. 28–32. ISSN: 1525-5050, 1525-5069. DOI: 10.1016/j.yebeh.2019.03.045. URL: https://www.epilepsybehavior.com/article/S1525-5050(19)30011-3/abstract (visited on 03/09/2024) (cit. on p. 7).

[25] Ingrid E. Scheffer et al. "ILAE classification of the epilepsies: Position paper of the ILAE Commission for Classification and Terminology". en. In: *Epilepsia* 58.4 (Apr. 2017), pp. 512–521. ISSN: 0013-9580, 1528-1167. DOI: 10.1111/epi.13709. URL: https://onlinelibrary.wiley.com/doi/10.1111/epi.13709 (visited on 03/25/2025) (cit. on pp. 8, 9).

[26] *EpilepsyDiagnosis.org*. URL: https://epilepsydiagnosis.org/index.html (visited on 03/09/2024) (cit. on p. 9).

[27] Sharika Raga et al. "Electroclinical markers to differentiate between focal and generalized epilepsies". en. In: *Epileptic Disorders* 23.3 (June 2021), pp. 437–458. ISSN: 1294-9361, 1950-6945. DOI: 10.1684/epd.2021.1291. URL: https://onlinelibrary.wiley.com/doi/10.1684/epd.2021.1291 (visited on 03/21/2025) (cit. on p. 9).

[28] Robert S. Fisher et al. "Operational classification of seizure types by the International League Against Epilepsy: Position Paper of the ILAE Commission for Classification and Terminology". en. In: *Epilepsia* 58.4 (Apr. 2017), pp. 522–530. ISSN: 0013-9580, 1528-1167. DOI: 10.1111/epi.13670. URL: https://onlinelibrary.wiley.com/doi/10.1111/epi.13670 (visited on 03/24/2025) (cit. on p. 9).

[29] M. Le Van Quyen et al. "Spatio-temporal characterizations of non-linear changes in intracranial activities prior to human temporal lobe seizures". en. In: *European Journal of Neuroscience* 12.6 (2000). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1046/j.1460-

9568.2000.00088.x, pp. 2124–2134. ISSN: 1460-9568. DOI: 10.1046/j.1460-9568.2000.00088.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1046/j.1460-9568.2000.00088.x (visited on 03/25/2025) (cit. on p. 11).

[30] Marco Mula. "Epilepsy-induced behavioral changes during the ictal phase". en. In: *Epilepsy & Behavior* 30 (Jan. 2014), pp. 14–16. ISSN: 15255050. DOI: 10.1016/j.yebeh.2013.09.011. URL: https://linkinghub.elsevier.com/retrieve/pii/S1525505013004770 (visited on 03/25/2025) (cit. on p. 11).

[31] Robert S. Fisher, Helen E. Scharfman, and Marco deCurtis. "How Can We Identify Ictal and Interictal Abnormal Activity?" en. In: *Issues in Clinical Epileptology: A View from the Bench*. Ed. by Helen E. Scharfman and Paul S. Buckmaster. Dordrecht: Springer Netherlands, 2014, pp. 3–23. ISBN: 978-94-017-8914-1. DOI: 10.1007/978-94-017-8914-1_1. URL: https://doi.org/10.1007/978-94-017-8914-1_1 (visited on 03/25/2025) (cit. on p. 11).

[32] Robert S. Fisher and Jerome J. Engel. "Definition of the postictal state: When does it start and end?" In: *Epilepsy & Behavior*. The Post-ictal State 19.2 (Oct. 2010), pp. 100–104. ISSN: 1525-5050. DOI: 10.1016/j.yebeh.2010.06.038. URL: https://www.sciencedirect.com/science/article/pii/S1525505010004592 (visited on 03/25/2025) (cit. on p. 11).

[33] Evren Burakgazi and Jacqueline A. French. "Treatment of epilepsy in adults". eng. In: *Epileptic Disorders: International Epilepsy Journal with Videotape* 18.3 (Sept. 2016), pp. 228–239. ISSN: 1950-6945. DOI: 10.1684/epd.2016.0836 (cit. on p. 12).

[34] Mika Shirasu and Kazushige Touhara. "The scent of disease: volatile organic compounds of the human body related to disease and disorder". In: *The Journal of Biochemistry* 150.3 (Sept. 2011), pp. 257–266. ISSN: 0021-924X. DOI: 10.1093/jb/mvr090. URL: https://doi.org/10.1093/jb/mvr090 (visited on 03/25/2025) (cit. on p. 12).

[35] F. Tassi et al. "Volatile organic compounds (VOCs) in air from Nisyros Island (Dodecanese Archipelago, Greece): Natural versus anthropogenic sources". eng. In: *Environmental Pollution (Barking, Essex: 1987)* 180 (Sept. 2013), pp. 111–121. ISSN: 1873-6424. DOI: 10.1016/j.envpol.2013.05.023 (cit. on p. 12).

[36] Robert S. Blake, Paul S. Monks, and Andrew M. Ellis. "Proton-Transfer Reaction Mass Spectrometry". In: *Chemical Reviews* 109.3 (Mar. 2009). Publisher: American Chemical Society, pp. 861–896. ISSN: 0009-2665. DOI: 10.1021/cr800364q. URL: https://doi.org/10.1021/cr800364q (visited on 03/25/2025) (cit. on p. 12).

[37] Hiroyuki Kataoka et al. "Noninvasive analysis of volatile biomarkers in human emanations for health and early disease diagnosis". eng. In: *Bioanalysis* 5.11 (June 2013), pp. 1443–1459. ISSN: 1757-6199. DOI: 10.4155/bio.13.85 (cit. on p. 13).

[38] L. Sahu. "Volatile organic compounds and their measurements in the troposphere". In: 2012. URL: https://www.semanticscholar.org/paper/Volatile-organic-compounds-and-their-measurements-Sahu/c06092fb6e315b6c7a33fccae7ad2d3c23646026 (visited on 03/26/2025) (cit. on p. 13).

[39] Vaughan S. Langford, Ian Graves, and Murray J. McEwan. "Rapid monitoring of volatile organic compounds: a comparison between gas chromatography/mass spectrometry and selected ion flow tube mass spectrometry". eng. In: *Rapid communications in mass spectrometry: RCM* 28.1 (Jan. 2014), pp. 10–18. ISSN: 1097-0231. DOI: 10.1002/rcm.6747 (cit. on p. 13).

[40] Ruben Epping and Matthias Koch. "On-Site Detection of Volatile Organic Compounds (VOCs)". In: *Molecules* 28.4 (2023). ISSN: 1420-3049. DOI: 10.3390/molecules28041598. URL: https://www.mdpi.com/1420-3049/28/4/1598 (cit. on p. 13).

[41] Revathi Rajan et al. "Chemical Fingerprinting of Human Body Odor: An Overview of Previous Studies". In: *Malaysian Journal of Forensic Science* (Apr. 2014) (cit. on p. 13).

[42] Nijing Wang et al. "Emission Rates of Volatile Organic Compounds from Humans". In: *Environmental Science & Technology* 56.8 (Apr. 2022). Publisher: American Chemical Society, pp. 4838–4848. ISSN: 0013-936X. DOI: 10.1021/acs.est.1c08764. URL: https://doi.org/10.1021/acs.est.1c08764 (visited on 03/26/2025) (cit. on p. 13).

[43] L Blanchet et al. "Factors that influence the volatile organic compound content in human breath". en. In: *Journal of Breath Research* 11.1 (Feb. 2017). Publisher: IOP Publishing, p. 016013. ISSN: 1752-7163. DOI: 10.1088/1752-7163/aa5cc5. URL: https://dx.doi.org/10.1088/1752-7163/aa5cc5 (visited on 03/26/2025) (cit. on p. 13).

[44] Xiaochen Tang et al. "Volatile Organic Compound Emissions from Humans Indoors". eng. In: *Environmental Science & Technology* 50.23 (Dec. 2016), pp. 12686–12694. ISSN: 1520-5851. DOI: 10.1021/acs.est.6b04415 (cit. on pp. 13, 14).

[45] Daniel Gaisberger. *An analysis of the efficiency of machine learning algorithms in the detection and prediction of epileptic seizures*. eng. 2024. URL: https://resolver.obvsg.at/urn:nbn:at:at-ubl:1-73065 (cit. on pp. 15, 16).

[46] Dieuwke van Dartel et al. "Breath analysis in detecting epilepsy". en. In: *Journal of Breath Research* 14.3 (Apr. 2020). Publisher: IOP Publishing, p. 031001. ISSN: 1752-7163. DOI: 10.1088/1752-7163/ab6f14. URL: https://dx.doi.org/10.1088/1752-7163/ab6f14 (visited on 11/18/2024) (cit. on p. 16).

[47] Philip Davis. "The Investigation of Human Scent from Epileptic Patients for the Identification of a Biomarker for Epileptic Seizures". English. ISBN: 9780438323292 Publication Title: ProQuest Dissertations and Theses. Ph.D. United States – Florida: Florida International University, 2017. URL: https://www.proquest.com/docview/2103114969/abstract/E7FD2BF3A85A4E09PQ/1 (visited on 11/18/2024) (cit. on p. 16).

[48] E.I. Mohamed et al. "Machine Learning-Based Electronic Nose for Universal Mapping of Blood Odors and Diagnosis of Cancer". In: *Chemical Engineering Transactions* 112 (2024), pp. 121–126. DOI: 10.3303/CET24112021 (cit. on p. 17).

[49] V.A. Binson, M. Subramoniam, and L. Mathew. "Prediction of lung cancer with a sensor array based e-nose system using machine learning methods". In: *Microsystem Technologies* 30.11 (2024), pp. 1421–1434. DOI: 10.1007/s00542-024-05656-5 (cit. on p. 17).

[50] Q. Wang et al. "Diagnostic performance of volatile organic compounds analysis and electronic noses for detecting colorectal cancer: a systematic review and meta-analysis". In: *Frontiers in Oncology* 14 (2024). DOI: 10.3389/fonc.2024.1397259 (cit. on p. 17).

[51] E.I. Mohamed et al. "Volatile organic compounds of biofluids for detecting lung cancer by an electronic nose based on artificial neural network". In: *Journal of Applied Biomedicine* 17.1 (2019), pp. 61–67. DOI: 10.32725/jab.2018.006 (cit. on p. 17).

[52] R.E. Evenhuis et al. "Diagnosis of chondrosarcoma in a noninvasive way using volatile organic compounds in exhaled breath: a pilot study". In: *Future Oncology* 20.22 (2024), pp. 1545–1552. DOI: 10.1080/14796694.2024.2355080 (cit. on p. 17).

[53] T. Sukaram et al. "Exhaled volatile organic compounds for diagnosis of hepatocellular carcinoma". In: *Scientific Reports* 12.1 (2022). DOI: 10.1038/s41598-022-08678-z (cit. on p. 17).

[54] M.L. Bastos et al. "Breaking barriers in Candida spp. detection with Electronic Noses and artificial intelligence". In: *Scientific Reports* 14.1 (2024). DOI: 10.1038/s41598-023-50332-9 (cit. on p. 17).

[55] G. Petkov et al. "Electroencephalographic events prior to epileptic major motor seizures". In: *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (Aug. 2012). Conference Name: 2012 34th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) ISBN: 9781457717871 9781424441198 Place: San Diego, CA Publisher: IEEE, pp. 1028–1031. DOI: 10.1109/EMBC.2012.6346109. URL: http://ieeexplore.ieee.org/document/6346109/ (visited on 11/18/2024) (cit. on p. 25).

[56] D M Murray et al. "Defining the gap between electrographic seizure burden, clinical expression and staff recognition of neonatal seizures". en. In: *Archives of Disease in Childhood - Fetal and Neonatal Edition* 93.3 (May 2008), F187–F191. ISSN: 1359-2998, 1468-2052. DOI: 10.1136/adc.2005.086314. URL: https://fn.bmj.com/lookup/doi/10.1136/adc.2005.086314 (visited on 11/18/2024) (cit. on p. 25).

[57] Petr Kriz et al. "Unveiling the Smell Inspector and Machine Learning Methods for Smell Recognition". In: *2023 15th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT)*. ISSN: 2157-023X. Oct. 2023, pp. 182–187. DOI: 10.1109/ICUMT61075.2023.10333105. URL: https://ieeexplore.ieee.org/document/10333105 (visited on 12/11/2024) (cit. on p. 36).

[58] Ruud Peters et al. "Evaluation of a Commercial Electronic Nose Based on Carbon Nanotube Chemiresistors". en. In: *Sensors* 23.11 (Jan. 2023). Number: 11 Publisher: Multidisciplinary Digital Publishing Institute, p. 5302. ISSN: 1424-8220. DOI: 10.3390/s23115302. URL: https://www.mdpi.com/1424-8220/23/11/5302 (visited on 12/11/2024) (cit. on p. 36).

[59] Jason Poulos and Rafael Valle. "Missing Data Imputation for Supervised Learning". In: *Applied Artificial Intelligence* 32.2 (Apr. 2018). Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/08839514.2018.1448143, pp. 186–196. ISSN: 0883-9514. DOI: 10.1080/08839514.2018.1448143. URL: https://doi.org/10.1080/08839514.2018.1448143 (visited on 01/13/2025) (cit. on p. 45).

[60] Lean Yu et al. "Missing Data Preprocessing in Credit Classification: One-Hot Encoding or Imputation?" In: *Emerging Markets Finance and Trade* 58.2 (Jan. 2022). Publisher: Routledge _eprint: https://doi.org/10.1080/1540496X.2020.1825935, pp. 472–482.

ISSN: 1540-496X. DOI: 10.1080/1540496X.2020.1825935. URL: https://doi.org/10.1080/1540496X.2020.1825935 (visited on 01/13/2025) (cit. on p. 46).

[61] Aishwarya Asesh. "Normalization and Bias in Time Series Data". en. In: ed. by Cezary Biele et al. Vol. 440. Book Title: Digital Interaction and Machine Intelligence Series Title: Lecture Notes in Networks and Systems. Cham: Springer International Publishing, 2022, pp. 88–97. ISBN: 978-3-031-11431-1 978-3-031-11432-8. DOI: 10.1007/978-3-031-11432-8_8. URL: https://link.springer.com/10.1007/978-3-031-11432-8_8 (visited on 01/20/2025) (cit. on p. 47).

[62] M. Löning et al. "sktime: A Unified Interface for Machine Learning with Time Series". In: *ArXiv* (Sept. 2019). URL: https://www.semanticscholar.org/paper/sktime%3A-A-Unified-Interface-for-Machine-Learning-L%C3%B6ning-Bagnall/38eb43ea485e7d288c75ee514e0d51b8bffa3d18 (visited on 11/29/2024) (cit. on p. 48).

[63] Wes McKinney. "Data Structures for Statistical Computing in Python". In: *Proceedings of the 9th Python in Science Conference*. Ed. by Stéfan van der Walt and Jarrod Millman. 2010, pp. 56–61. DOI: 10.25080/Majora-92bf1922-00a (cit. on p. 48).

[64] Charles R. Harris et al. "Array programming with NumPy". In: *Nature* 585.7825 (Sept. 2020). Publisher: Springer Science and Business Media LLC, pp. 357–362. DOI: 10.1038/s41586-020-2649-2. URL: https://doi.org/10.1038/s41586-020-2649-2 (cit. on p. 48).

[65] Mojtaba Bandarabadi et al. "On the proper selection of preictal period for seizure prediction". In: *Epilepsy & Behavior* 46 (May 2015), pp. 158–166. ISSN: 1525-5050. DOI: 10.1016/j.yebeh.2015.03.010. URL: https://www.sciencedirect.com/science/article/pii/S1525505015001158 (visited on 12/23/2024) (cit. on p. 52).

[66] Syed Muhammad Usman, Muhammad Usman, and Simon Fong. "Epileptic Seizures Prediction Using Machine Learning Methods". en. In: *Computational and Mathematical Methods in Medicine* 2017 (2017), pp. 1–10. ISSN: 1748-670X, 1748-6718. DOI: 10.1155/2017/9074759. URL: https://www.hindawi.com/journals/cmmm/2017/9074759/ (visited on 12/23/2024) (cit. on p. 53).

[67] Rajlakshmi Borthakur et al. "Framework to select the top ML classifier for robust seizure detection and prediction: A comparison-based study using multiple preictal time and feature sets". In: Repository: In Review. Apr. 2021. DOI: 10.21203/rs.3.rs-

470605/v1. URL: https://www.researchsquare.com/article/rs-470605/v1 (visited on 12/23/2024) (cit. on p. 53).

[68] Nicolas Zurbuchen, Adriana Wilde, and Pascal Bruegger. "A Machine Learning Multi-Class Approach for Fall Detection Systems Based on Wearable Sensors with a Study on Sampling Rates Selection". en. In: *Sensors* 21.3 (Jan. 2021), p. 938. ISSN: 1424-8220. DOI: 10.3390/s21030938. URL: https://www.mdpi.com/1424-8220/21/3/938 (visited on 12/23/2024) (cit. on p. 53).

[69] Scott R Small et al. "Impact of Reduced Sampling Rate on Accelerometer-based Physical Activity Monitoring and Machine Learning Activity Classification". en. In: (Oct. 2020). Repository: Public and Global Health. DOI: 10.1101/2020.10.22.20217927. URL: http://medrxiv.org/lookup/doi/10.1101/2020.10.22.20217927 (visited on 12/23/2024) (cit. on p. 53).

[70] Qiong Wei and Roland L. Dunbrack. "The Role of Balanced Training and Testing Data Sets for Binary Classifiers in Bioinformatics". en. In: *PLoS ONE* 8.7 (July 2013). Ed. by Iddo Friedberg, e67863. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0067863. URL: https://dx.plos.org/10.1371/journal.pone.0067863 (visited on 12/21/2024) (cit. on p. 54).

[71] Jinseok Kim and Jenna Kim. "The impact of imbalanced training data on machine learning for author name disambiguation". en. In: *Scientometrics* 117.1 (Oct. 2018), pp. 511–526. ISSN: 0138-9130, 1588-2861. DOI: 10.1007/s11192-018-2865-9. URL: http://link.springer.com/10.1007/s11192-018-2865-9 (visited on 12/21/2024) (cit. on p. 54).

[72] Siddharth Pramod et al. "Detecting Epileptic Seizures from EEG Data using Neural Networks". In: *ArXiv* (Dec. 2014). URL: https://www.semanticscholar.org/paper/Detecting-Epileptic-Seizures-from-EEG-Data-using-Pramod-Page/9c924a22c7688cf80fe2f7cb28862d901ce5eb6d (visited on 12/25/2024) (cit. on p. 57).

[73] Mustafa Umit Oner et al. "Training machine learning models on patient level data segregation is crucial in practical clinical applications". en. In: (Apr. 2020). Repository: Health Informatics. DOI: 10.1101/2020.04.23.20076406. URL: http://medrxiv.org/lookup/doi/10.1101/2020.04.23.20076406 (visited on 12/23/2024) (cit. on p. 57).

[74] Sina Shafiezadeh et al. "Methodological Issues in Evaluating Machine Learning Models for EEG Seizure Prediction: Good Cross-Validation Accuracy Does Not Guarantee Generalization to New Patients". en. In: *Applied Sciences* 13.7 (Mar. 2023), p. 4262. ISSN: 2076-3417. DOI: 10.3390/app13074262. URL: https://www.mdpi.com/2076-3417/13/7/4262 (visited on 12/25/2024) (cit. on p. 57).

[75] Christoph Bergmeir and José M. Benítez. "On the use of cross-validation for time series predictor evaluation". en. In: *Information Sciences* 191 (May 2012), pp. 192–213. ISSN: 00200255. DOI: 10.1016/j.ins.2011.12.028. URL: https://linkinghub.elsevier.com/retrieve/pii/S0020025511006773 (visited on 12/25/2024) (cit. on p. 57).

[76] Anita Rácz, Dávid Bajusz, and Károly Héberger. "Multi-Level Comparison of Machine Learning Classifiers and Their Performance Metrics". en. In: *Molecules* 24.15 (Aug. 2019), p. 2811. ISSN: 1420-3049. DOI: 10.3390/molecules24152811. URL: https://www.mdpi.com/1420-3049/24/15/2811 (visited on 12/01/2024) (cit. on p. 59).

[77] Gireen Naidu, Tranos Zuva, and Elias Mmbongeni Sibanda. "A Review of Evaluation Metrics in Machine Learning Algorithms". en. In: *Artificial Intelligence Application in Networks and Systems*. Ed. by Radek Silhavy and Petr Silhavy. Cham: Springer International Publishing, 2023, pp. 15–25. ISBN: 978-3-031-35314-7. DOI: 10.1007/978-3-031-35314-7_2 (cit. on p. 60).

[78] Bradley J. Erickson and Felipe Kitamura. "Magician's Corner: 9. Performance Metrics for Machine Learning Models". In: *Radiology: Artificial Intelligence* 3.3 (May 2021). Publisher: Radiological Society of North America, e200126. DOI: 10.1148/ryai.2021200126. URL: https://pubs.rsna.org/doi/10.1148/ryai.2021200126 (visited on 12/09/2024) (cit. on p. 60).

[79] Indranil Balki et al. "Sample-Size Determination Methodologies for Machine Learning in Medical Imaging Research: A Systematic Review". In: *Canadian Association of Radiologists Journal* 70.4 (Nov. 2019), pp. 344–353. ISSN: 0846-5371. DOI: 10.1016/j.carj.2019.06.002. URL: https://www.sciencedirect.com/science/article/pii/S0846537119300506 (visited on 01/07/2025) (cit. on p. 84).

[80] Ahmed H. Jalal et al. "Prospects and Challenges of Volatile Organic Compound Sensors in Human Healthcare". eng. In: *ACS sensors* 3.7 (July 2018), pp. 1246–1263. ISSN: 2379-3694. DOI: 10.1021/acssensors.8b00400 (cit. on p. 84).

[81] Claire Turner. "Techniques and issues in breath and clinical sample headspace analysis for disease diagnosis". eng. In: *Bioanalysis* 8.7 (Apr. 2016), pp. 677–690. ISSN: 1757-6199. DOI: 10.4155/bio.16.22 (cit. on p. 84).

[82] Kristin Schallschmidt et al. "Comparison of volatile organic compounds from lung cancer patients and healthy controls—challenges and limitations of an observational study". en. In: *Journal of Breath Research* 10.4 (Oct. 2016). Publisher: IOP Publishing, p. 046007. ISSN: 1752-7163. DOI: 10.1088/1752-7155/10/4/046007. URL: https://dx.doi.org/10.1088/1752-7155/10/4/046007 (visited on 01/07/2025) (cit. on p. 85).