800 ANNI 1222·2022

UNIVERSITÀ DEGLI STUDI DI PADOVA

DEI DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

UNIVERSITY OF PADOVA

Department of Information Engineering

Master Degree in ICT for Internet and Multimedia

Attention-based EEG classification

SUPERVISOR
Prof. Giulia Cisotto

CANDIDATE
Laura Soccol

Date: 05/12/2022

A.Y.: 2021/2022

Acknowledgements

I would like to thank a lot of people, but that would take too long.

I would like to express my sincere gratitude to my thesis supervisor Prof. Giulia Cisotto for her assistance at every stage of this thesis. I would also like to thank all the research project members, especially my internship tutor Prof. Urszula Markowska-Kaczmar for her comments and suggestions.

I would like to thank my family and my sister for their constant support.

Last but not least, I would like to thank all my friends who have shared with me these years.

Contents

Abstract

Electroencephalography signal is an important resource for the analysis of the brain. However, due to its complexity and non-stationarity, its analysis has often been a challenging task for physicians. Machine learning and deep learning have shown attractive results in different scenarios, and they could be effective means to help clinicians to diagnose abnormal conditions and to speed up the diagnosis process. However, the complexity of these models prevents from fully understanding what really happens inside them, and the way input information is used. Attention, one of the most recent developments in the DL field, allows the models to learn which input information is useful to perform classification. In this thesis, we compare two commonly used DL models, namely CNN and LSTM, with their attention-enhanced counterparts. These models are tested on three different datasets, related to three different challenges in EEG research, which are abnormalities detection, artifact detection, and seizure classification. We achieved the state of the art in all classification problems, regardless of the large variability of the datasets and the simple employed architecture. Moreover, the use of attention provides an increase in the final accuracy, highlighting a promising strategy to identify the relevant information in the EEG signal.

Introduction

Electroencephalography (EEG) has been widely used through the years to detect abnormalities in the brain [1]. However, the evaluation of EEG traces by clinicians is often a time-consuming process, and it requires years of training and experience to recognize pathological patterns. Moreover, the diagnosis accuracy depends on the experts' training and experience [2]. The use of machine learning (ML) and deep learning (DL) methods allows to extract features and detect patterns that can't be recognized by humans, enhancing the accuracy of clinicians' work. Moreover, these techniques can learn a new task in just a few hours, making diagnosis faster.

Starting from the simplest and oldest neural network (NN) model [3], which was inspired by the brain structure, several other advanced architectures have been proposed through the years, making the learning process more and more efficient. Convolutional Neural Networks (CNN) [4], inspired by the visual cortex, take into account the idea of spatial correlation. Then, Recurrent Neural Network (RNN) implements the idea of predictive processing where each current decision is driven by past information, modeling the temporal dependencies [5]. Finally, the attention mechanism aims to focus on certain inputs, ignoring the rest [6]. All these models are used in the analysis of complex signals such as EEG leading to very impressive results.

In this thesis we compare two commonly used DL models, namely CNN and LSTM, with and without the attention mechanisms enhancement, to evaluate their impact on the classification performances.

Three different datasets are selected from the TUH EEG corpus which are related to three different challenges in EEG research. The TUH Abnormal EEG Corpus (TUAB) [7], contains generic abnormal EEG trials and normal EEG traces, the TUH Artifact EEG Corpus (TUAR) [8] provides clean EEG signals, and signals corrupted by 5 different kinds of artifacts, the TUH Seizure EEG Corpus (TUSZ) [9] consists of EEG trials affected by different types of epileptic seizures. For all datasets a binary classification problem is approached, thus we test the ability of the network to correctly classify abnormal and normal data in the first case, signals affected by artifact and clean segments in the second, and to discriminate among focal and global seizure in the last.

The classification performances are measured through the commonly used metrics: accuracy, recall, precision and F1-score.
The attention mechanism leads to an increase in the average accuracy in almost all the datasets for both models. In the CNN-based one, the accuracy increased from 71.41% to 74.24% in the TUAB dataset, from 84.36% to 87.83% in the TUAR, and from 84.96% to 86.92% in the TUSZ. Instead, the LSTM-based network accuracy increases from 72.94% to 74.03% in the TUAB dataset, and from 87.52% to 89.36% in the TUAR. The TUSZ dataset is the least affected by the attention mechanism improvement, and the final accuracy slightly overtakes that obtained in the simple LSTM (without attention) from 88.11% to 88.22%.

The rest of the thesis is organized as follows. Chapter 1 introduces the theoretical background on which the thesis is based. First, it describes how EEG signal originates and propagates, how it is recorded, and the most common preprocessing step. Then the traditional ML algorithms and the more advanced DL models are introduced. Finally, a literature review of the most recent attention mechanisms implementation is presented. In Chapter 2 the used datasets are introduced, together with all the processing steps, i.e. the preprocessing pipeline, the features extraction, and the features selection. Finally, the used models, the training pipeline, and the classification metrics are described. Chapter 3 is dedicated to results and discussion. The thesis is concluded by highlighting the main take-home messages of this work and the introduction of some promising future perspectives.

Chapter 1

Background

1.1 Brain physiology

1.1.1 The nervous system

The nervous system is a complex structure of nerves, cells, and organs that allow messages transmission around the brain and the rest of the body [10]. Signals from the outside are captured by specialized sense receptors, encoded, and transmitted to the brain through nerves. Here, the stimulus is processed, the appropriate response is produced, and transmitted back to control muscles, glands, and organs. The nervous system is divided into two parts: the Central Nervous System (CNS) which is made up of the brain and the spinal cord, and the Peripheral Nervous System (PNS) consisting of all the nerves transmitting the response signal from the brain to all the rest of the body. PNS could be in turn divided into two parts: the somatic nervous system, which acquires information from the sensory systems and controls the voluntary body movements, and the autonomic nervous system which governs the unconscious activity of the organs and glands. The core of the CNS, and also the most complex part of the human body is the brain. Encased in the skull and protected by the meninges, it is divided into three parts: the cerebrum, the cerebellum, and the brainstem [11]. The cerebrum is the most important

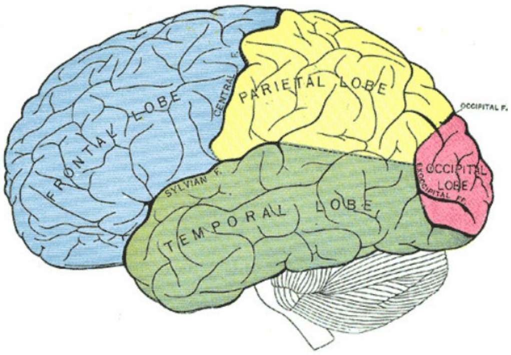


Figure 1.1: Lateral view of the cerebrum lobes.

part. It is split into two hemispheres, connected through a bundle of nerve fibers which ensures the communication between the two areas. Although the two parts look symmetrical, they are dominant in the control of different functions. Two different kinds of tissues are found in the brain. The gray matter is the external part, that is the cerebral cortex. It is mainly composed of neurons and glia cells and it is involved in information processing tasks. Then, the white matter is mainly composed of axons and it is responsible for transmitting signals between the gray matter and the rest of the nervous system. The cortex surface is characterized by many folds which allow increasing the effective surface area of the grey matter, and consequently the amount of information that could be processed. Each hemisphere is divided into four *lobes* as shown in Figure 1.1.

Each *lobe* has a different function [12]:

- **frontal lobe:** located in the forward part, at the front of the central sulcus, it is involved in motor functions, problem solving, reasoning and planning, language and emotion regulation;

- **parietal lobe:** it is just behind the frontal lobe, and it is responsible of processes information coming from the different parts of the body;

- **temporal lobe:** located at the bottom of the brain, under the lateral fissure. It is the location of the primary auditory cortex, thus its main function is to process auditory sounds. It is also involved with memory and new information processing;

- **occipital lobe:** it is in the back of the brain and it is employed in processing visual information.

1.1.2 Signal generation

The brain can be seen as a complex network whose fundamental unit is the neuron. A human brain consists of 10^{11} neurons, with 10^4 neurons mm^3 on average. Neurons are specialized cells made up by three parts as shown in Figure 1.8: the body (soma), which is the central part that contains the nucleus, the dendrites, branched ramification that receives signals from other neurons, and propagate toward the soma, and the axon, a single, tail-like ramification that transmits signals toward other neurons through the axon terminal. Axons are coated by a fatty substance called *myelin* that helps axons to conduct an electrical signal. At the end of the axon, the synapses manage the signal transmission from one neuron to another.

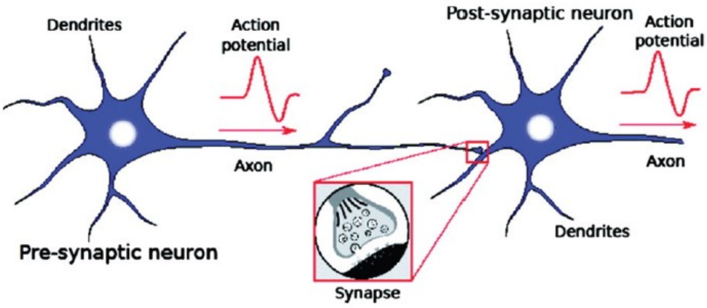


Figure 1.2: The neuron and the action potential transmission [13]

Information is transmitted as *action potentials*, which are temporary changes in the polarization of the cell membrane. The latter acts as a capacitor, separating the electrically charged particles of the external environment from the ions inside the cell. In normal conditions,

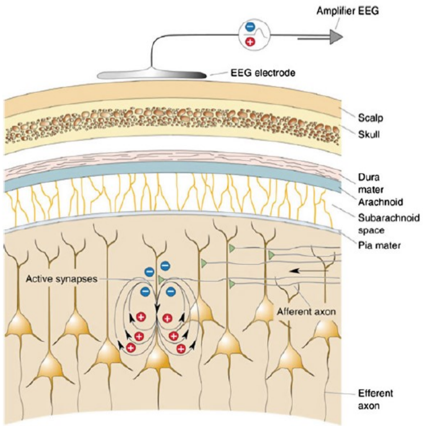


Figure 1.3: The signal propagation [15].

the resting state voltage is around -70 mV. When a stimulus comes and it exceeds a critical threshold (typically 15 mV), the membrane depolarizes very quickly [14]. Then, the action potential is propagated through the axon via local currents which induce depolarization of the adjacent part of the axon membrane. Transmission between two neurons instead, occurs through junctions called synapses. When the signal arrives at the axon terminal, it depolarizes the presynaptic membrane activating the stored neurotransmitters, both chemical or electrical. In turn, the postsynaptic membrane contains receptors for neurotransmitters. When they bind the receptors, a new action potential propagates in the postsynaptic cell.

1.1.3 Signal propagation

The burst of a single neuron can't be detected without a direct contact with it. However, when specific information has to be processed or transmitted, several neurons in the same area activate synchronously. Thus, the summation of all the inhibitory or excitatory postsynaptic potentials from pyramidal neurons, generate the electrical potentials which, passing through tissues, bone, and skull, can be measured from the head surface [15]. Since the signal has to flow through many different layers, the recording can be interpreted as the superposition of several components [16]:

$$EEG_{signal} = extLFP + propagationeffects + artifact + noise$$

where extLFP is the signal of interest, produced by the neurons, propagation effects are all the modifications introduced on the signal and related to the transmission along all the layers, and artifact and noise are other sources independent from the signal of interest. In the same way, the signal power suffers from the same attenuation effects.

1.2 The electroencephalogram

1.2.1 Signal characterization

The EEG signal is characterized by an oscillating behavior with an amplitude in the order of μV. It changes with time and it is strictly related to brain activity. A signal with high amplitude and low frequency represents the coordinated activity of a large brain area, while a low amplitude and high frequency stand for a desynchronized activity of neurons involved in different tasks [17]. On frequency content, brain activity covers the range [0.5-500] Hz. However, EEG can capture frequencies in the range [1.5-80] Hz, while for lower or higher frequencies techniques like Magnetoencephalography (MEG), Functional magnetic resonance imaging (fMRI), Electrocorticography (ECoG), or Local Field Potentials (LFP) could be used [18]. The five brain rhythms are [19]:

- **Delta [0.1-4] Hz:** it corresponds to the lowest possible frequencies but it is characterized by the highest amplitude. They constitute the dominant rhythm in infants and during sleep (especially stages 3 and 4), while they are abnormal in awake adults.

- **Theta [4-8] Hz:** is related to subconscious activity, deep relaxation, and meditation. Common in children, it is abnormal in

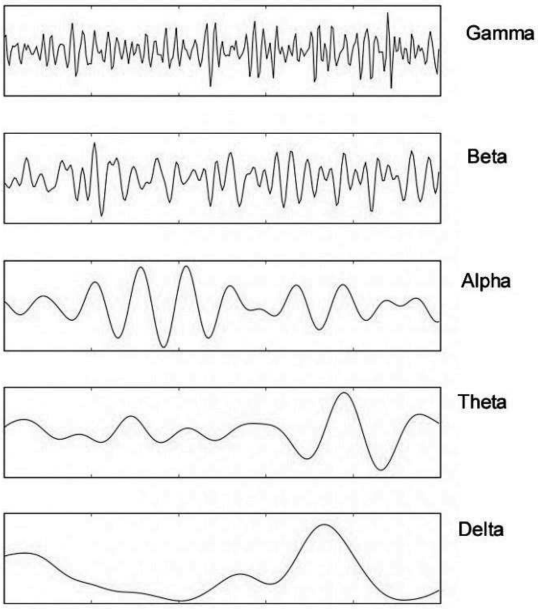


Figure 1.4: The fundamental EEG waves.

waking adults.

- **Alpha [8-12] Hz:** it is commonly found in healthy adults in relaxing conditions. It can be measured on both sides of the head, with a predominance on the nondominant side, and mainly on the occipital and parietal lobes.

- **Beta [12-30] Hz:** it occurs during the conscious state, and they are linked to action and concentration. It can be found in the frontal and parietal lobes on both sides.

- **Gamma [30-100] Hz:** it is related to simultaneous processing of information. An altered gamma activity can be related to many cognitive disorders.

1.2.2 Signal acquisition

EEG signal is acquired through electrodes placed on the scalp. Their location is fixed, and one of the most common acquisition models is given by the International 10-20 System (Figure 1.5). It provides

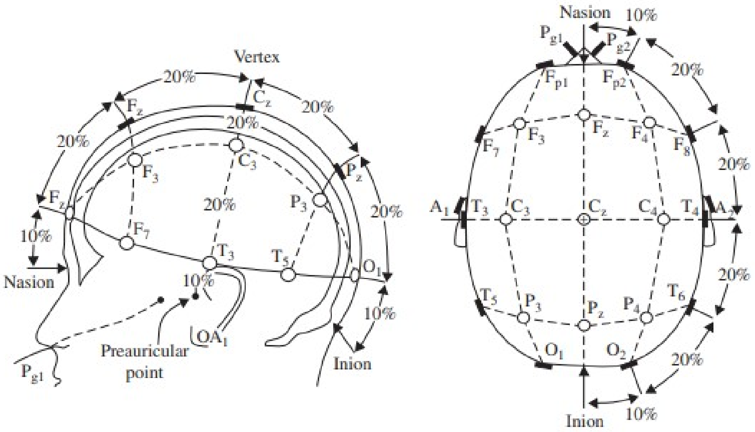


Figure 1.5: The International 10/20 system [21].

21 electrodes 2 of which act as references. Starting from some well-established points, namely nasion, inion, and peculiar points, the first electrodes are placed at a distance that is 10% the length of the scalp, while all the others are at a distance of 20% among them. Moreover, electrodes are marked with a letter and a number: the former derives from the lobe over which they are placed, and the latter identifies the position: odd electrodes are placed on the left hemisphere, while even ones on the right [20]. Three different kinds of modes can be used to record the signal [21]: differential, referential, or reference-free. The first measures the potential difference between pairs of electrodes (which are in input to each differential amplifier). The second computes the difference between an active electrode and a reference point. Reference electrodes could be Cz or others placed on the earlobes or mastoids. In the last mode, the reference comes from the average signal recorded by all the electrodes. Since the EEG signal amplitude is in the order of μV, it should be amplified. Finally, it is converted to digital with a sampling frequency at least of 250, and from 12 to 24 quantization bits. At this point, it is ready for further preprocessing steps like filtering, artifact removal or correction, and segmentation.

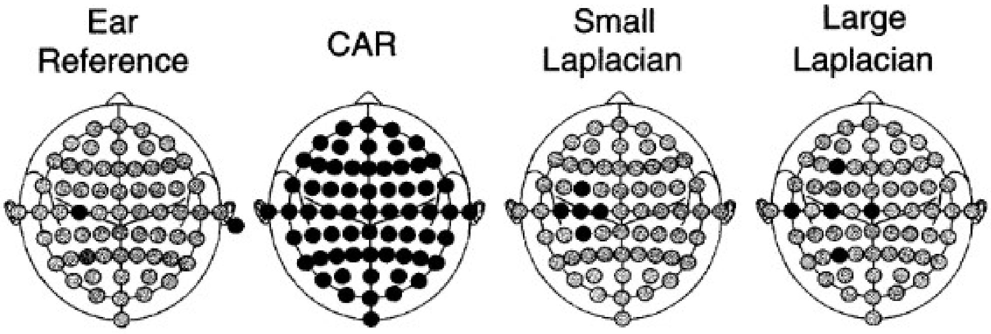


Figure 1.6: The spatial filters from [22].

1.2.3 Signal preprocessing

Preprocessing is an important step when working with EEG. It allows to clean raw data from artifacts or corrupted channels, to have a signal which contains for the most only useful information.

The first class of techniques is about filters. Through them, noise can be reduced or the SNR increased [22]. Two kinds of filters can be used: spatial and temporal ones. In the former (Figure 1.6), the amplitude of the signal coming from each electrode is changed according to the combination (weighted or not) of the voltages recorded at one or more (up to all) site locations. Four are the most commonly used techniques, that change with each other for the number or the location of the selected electrodes. In the ear reference technique, all channels are all referenced to an electrode placed out of the scalp, in general at the earlobe. In common average references (CAR), the average of the signal is subtracted from all the channels. In small Laplacian, for each electrode, the weighted sum of the four surrounding electrodes is considered, while in large Laplacian the same idea is followed but the used electrodes are those one electrode apart from the site of interest. On the other side, temporal filters allow to delete specific bands of frequency that don't contain signals of interest. A high pass filter can be used to delete baseline drifts, characterized by low frequency components, while a low pass one to remove electromagnetic interference or muscle contractions events. These two can also be combined in a band pass filter. Finally, a notch filter can be used to delete a specific frequency, like the power line noise at 50 Hz (or 60 Hz).

Segmentation is another preprocessing technique, which consists in dividing the whole signal into adjacent or overlapping segments. In this way, the EEG signal which is characterized by a non-stationary trend is segmented into chunks with similar time and frequency distributions. Moreover, portions corrupted by any kind of artifact can be removed without discarding the entire trace.

Once the signal is cleaned, informative features can be extracted. Many of them can be computed, like time domain, frequency domain, or time-frequency domain, then also more complex ones such as nonlinear features, entropies, or complex network [23]. In this thesis, the following features are used:

- **Mean:** it is the average value of all the N samples in the considered segment

$$\mu = \tfrac{1}{N}\sum_{i=0}^{N} x_i$$

- **Variance:** it measures the spread of the samples around mean

$$s^2 = \sum_{i=0}^{N} (x_i - \mu)^2$$

- **Zero crossing:** it is not a statistical feature, but it measures how many times the signal changes its sign.

$$zc = \sum_{i=0}^{N-1} 1_{\mathbb{R}<0(x_i,x_{i-1})}$$

- **Area under the curve (AUC):** is the integral of the EEG trace

$$AUC = \int_a^b |f(x)|\, dx$$

- **Peak to peak:** it measures the difference between the maximum and minimum amplitude found in the segment

$$p2p = max(x) - min(x)$$

- **Skewness:** it measures the asymmetry of the distribution. It can be positive, zero, negative, or undefined. In the case of unimodal distribution, negative skew indicates a left-sided tail, while a positive skew a right-sided tail.

$$skew = \frac{\sum_{i=0}^{N}(x_i-\mu)^3}{(N-1)\times\sigma^3}$$

- **Kurtosis:** it is another measure of shape and expresses the heaviness of a distribution's tails relative to a normal distribution.

$$kurt = \frac{\sum_{i=0}^{N}(x_i-\mu)^4}{(N-1)\times\sigma^4}$$

- **Spectral power:** it measures the frequency content of the signal. It is computed using Welch.

$$P_i = \frac{1}{\pi}\int_{w_1}^{w_2} S_x(w)\, dw$$

1.2.4 Clinical applications

EEG is widely used for clinical applications. It allows distinguishing between normal and abnormal acquisition, provides excellent temporal resolution to other techniques (in the order of ms), and could be used for continuous monitoring of patients. Finally, it is a non-invasive, and low-cost technique. The first field of application was epilepsy. In this case, EEG allows to identify the epileptic events, and the epileptogenic zone, and to asses if the therapy is good [24]. Other fields of application are the diagnosis of sleep disorder. Since each phase of sleep is characterized by a specific brain rhythm, EEG could be a useful tool in the diagnosis of sleep quality and related disorders [25]. It can also be used as a supplement to diagnose many different psychiatric disorders such as AHDH [26], depression [27], Alzheimer's disease [28], and schizophrenia [29]. Another emerging field of interest is brain computer interfaces (BCI). These devices could provide tetraplegic, post-stroke, or spinal-cord injured patients a way to communicate and interact with the outside world. In this case, the EEG signal could be translated into commands to control a machine and perform some actions [30].

1.3 Models for EEG analysis

EEG is an indispensable tool both for clinical purposes and to understand how it works [31]. However, it is often difficult to interpret, due to its high dimensionality, and non-stationarity. Moreover, its low SNR makes it difficult to discriminate between the signal and any other kind of noise or artifacts. Moreover, the evaluation of an EEG trace depends also on the training experience of the clinician [2]. In last years, machine learning (ML) and deep learning (DL) techniques have been introduced in the neuroscience field, providing an automated tool that could support clinicians in decoding the signal.

1.3.1 Traditional machine learning approaches

ML techniques rely on the extraction of handcrafted features. This means that some preprocessing steps should be performed on raw data. This means that different kinds of features have to be extracted and then fed in input to the classifier. However, this implies that researchers should have some a priori knowledge of which feature set is more suitable to investigate a specific task since the algorithm's final performances are strictly related to the selected features.

K-Nearest Neighbours (KNN)

K-Nearest Neighbours (KNN) is one of the simplest ML algorithms. It is nonparametric, thus it doesn't make any assumption on the mapping function between input and output data, but it just stores training data which will be used to classify new data points. It is based on the idea of similarity, which is that similar samples are near to each other, while samples originating by different distribution are in a different cluster. When a new data point is given as input, it is assigned to a class based on the similarity with the K nearest training data point around it. The classification function is based on the distance. An example is given by the following [32]:

$$w_{(i)} = \begin{cases} \frac{d_{(k)}-d_{(i)}}{d_{(k)}-d_{(1)}} if d_{(k)} \neq d_{(1)} \\ 1 if d_{(k)} = d_{(1)} \end{cases}$$

where d_1 is the distance between the input data and the nearest neighbor, d_k the distance to the further, and d_i is that of the i-th neighbor. This kind of model can be used to discriminate abnormal brain activity as in [33], to detect seizure precursors. The idea is that brain features coming from normal signals should be more similar to each other than features coming from abnormal signals, thus the use of KNN could help to associate a new EEG sample to one of the two classes. In another work [34], an adaptive KNN algorithm is used to detect the onset of epileptic events using discrete wavelet transform (DWT) as input features.

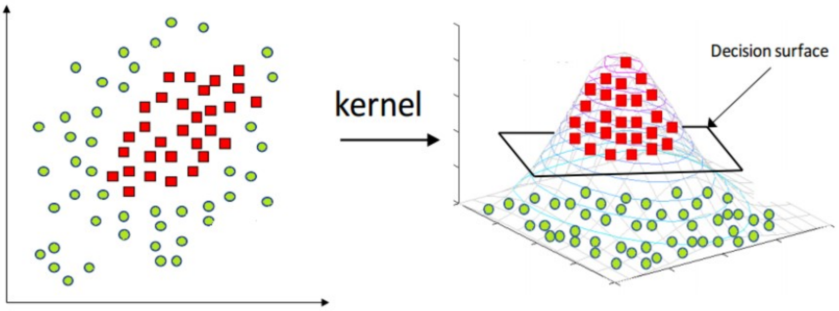


Figure 1.7: Example of SVM with the kernel trick.

Support Vector Machine (SVM)

Support Vector Machine (SVM) is a supervised algorithm that aims to find a separating hyperplane between samples of two classes, in such a way that the margin between the samples is maximized. The algorithm in its basic formulation works well for linearly separable data, while no good solution could be found in the case of not linearly separable data. To overcome this problem, data are mapped in a higher dimensional space, so that they become linearly separable (Figure 1.7). This step is crucial, and the choice of an incorrect kernel could lead to very poor outcomes. Then the decision boundaries are mapped back to the original feature dimension space. The learning process involves the minimization of a cost function, while the used loss is the Hinge loss which allows the maximum margin classification [35]:

$$l(x, y, f(x)) = max\{0, 1 - yf(x)\}$$

Despite the development of DL models, SVM is still widely used, due to its low complexity and adaptability in solving classification problem as brain disorder diagnosis [31], artifact [36] or seizure detection [37].

In [38] SVM is used to discriminate among signals coming from different cognitive conditions. Discrete wavelet-based features are extracted, normalized, and then selected to delete nonrelevant features.

Depending on the classification problem, the SVM algorithm resulted in an accuracy up to 99.11%. In [36] a novel approach to detect and remove artifacts is presented. Through the combined use of wavelet-ICA and SVM, artifacts have been removed with minimal distortion of the signal and without the need for visual inspection or manually defined thresholds. Results show an accuracy of 99.1%.

Linear discriminant analysis (LDA)

Linear discriminant analysis (LDA) is a simple and computationally efficient classifier. Similar to SVM, it was originally implemented to work with binary data and linear data. Moreover, if SVM tries to learn a hyperplane that divides data by maximizing the margin, LDA tries to divide data into two classes by maximizing the distance among different classes and minimizing the distance between points in the same class. LDA could be extended to multiclass problems, where multiple hyperplanes are learned. This algorithm has been implemented in BCI applications as in [39] and [40] leading to good results. However, the linearity of the model prevents competitive results on nonlinear EEG data [31].

Bayesian classifier

Bayesian classifier belongs to the probabilistic classifiers family. It aims to assign samples to the class to which they are more likely to belong, using Bayes's rule to compute the *a posterior* probability $P(y|x)$. For binary classification problems, that probability is computed as follows [41]:

$$(y|x) = \frac{P(y)P(x|y)}{P(x)} = \frac{P(y)P(x|y)}{P(A)P(x|A)+P(B)P(x|B)} = \frac{P(x|y)}{P(x|A)+P(x|B)}$$

where A and B are the two different classes and x is the vector representing the two independent variables. The *a priori* probability can be computed as a weighted mixture of Gaussian functions:

$$P(x|y) =_{i=01}^{N} w_iP(x|c_i)$$

where w_i is the weight of each Gaussian function and N the total number of them. The main advantage of this method is that, even with a small number of training points, it can successfully estimate the learning parameters, and reach an accuracy comparable with SVM ones [42]. However, it considers all feature vectors as independent regardless of any actual correlation. Bayes classifiers are successfully used in [38] for cognitive condition classification, while in [43] author used a novel parallel classifier, Parallel Genetic Naive Bayes Seizure Classifier (PGNBSZ) to classify the epileptic seizure.

1.3.2 Deep learning models

Artificial neural network (ANN) is a model which comes to the attention of researchers quite recently. Despite the first model was introduced back in '40 [3], the limitation in computational power and the lack of sufficiently large datasets for training prevent them to be used in practice.

The computation model is inspired by the structure of the brain network, and the learning process tries to mimic the generalization capabilities of the human brain. The basic computing unit of an ANN is the neuron (Figure 1.8. It receives information from input connections. Every connection is characterized by a weight that controls how much information should be considered. Then the computational units apply a nonlinear activation function to the weighted input. Then the output is propagated to the next unit.

An ANN is composed of several layers of these neurons, hierarchically organized, which receive as input the output of the previous layer and forward the processed output to the subsequent layer. Learning

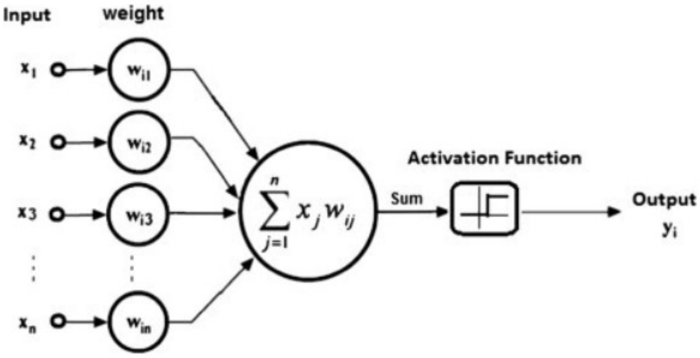
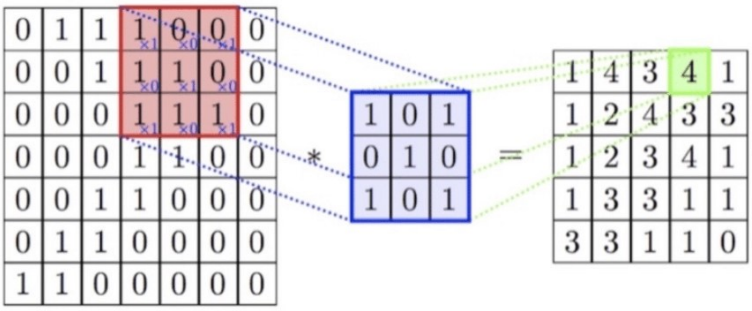


Figure 1.8: The basic ANN unit.

is the procedure through which the weights are iteratively adjusted to learn the function which maps the input into the output. At the end of the forward propagation, the loss is computed as a function of the learned weight. This loss is feeding backward (backward propagation) to fine-tune the weights. Optimization is then based on the Stochastic Gradient Descent (SGD) [44], an optimization algorithm that tries to minimize the error by moving in the opposite direction to the gradient of the loss function.

Neural networks are considered universal approximators. This means that every network with at least one hidden layer of nonlinear units can approximate any continuous function, provided the network has enough hidden units [45]. However, there is no guarantee that NN can learn the function since there is no theoretical proof of the SGD convergence [46]. Moreover, it could require an exponentially large number of units or computational time which is infeasible in practice. In the last years, ANNs have started to be used in medical applications [47]. Differently from the previously described ML models, they allow an end-to-end decoding procedure, which accepts as input raw or minimally processed data, avoiding the feature extraction process. The network itself can learn the informative features [2]. Moreover, the digitization of healthcare data provides more efficient data collection and sharing among multiple hospitals and research centers,



$$\begin{bmatrix} 0 & 1 & 1 & 1_{\times1} & 0_{\times0} & 0_{\times1} & 0 \\ 0 & 0 & 1 & 1_{\times0} & 1_{\times1} & 0_{\times0} & 0 \\ 0 & 0 & 0 & 1_{\times1} & 1_{\times0} & 1_{\times1} & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} * \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 4 & 3 & 4 & 1 \\ 1 & 2 & 4 & 3 & 3 \\ 1 & 2 & 3 & 4 & 1 \\ 1 & 3 & 3 & 1 & 1 \\ 3 & 3 & 1 & 1 & 0 \end{bmatrix}$$

Figure 1.9: The convolution between input and kernel.

allowing for huge databases for network training.

Convolutional neural network (CNN)

Convolutional Neural Network (CNN) architectures were proposed to solve some issues found in the NN models. Inspired by the visual cortex and the fact that cortical neurons respond to a stimulus only in a restricted portion of the whole visual field [48], they introduce the idea of local correlation on the NN, which means that closer input samples are more related than other farther ones. This was particularly relevant for image processing, where a pixel in an image is more related to the close ones than to those far apart, and CNN become one of the most used algorithms in computer vision. Moreover, connecting only neighbor neurons allows for reducing the number of connections and so the number of weights to tune.

Learning on CNN is based on convolution (Figure 1.9). Each convolution layer consists of a set of filters (or kernels), which are convolved with the input. Each kernel is a matrix of trainable weights that will be optimized during the training procedure. Different kernels generate different feature maps starting from the same input. In the lower layers, filters start encoding basic features, like for example edges, while proceeding deeper and deeper features increase their complexity to uniquely identify the input object. On the top of the last convolutional layer, a fully connected (dense) layer performs the classification.

At first used for image classification, it was then introduced in other areas, such as for time sequence classification. In [49] authors proposed the first CNN-based framework with transfer learning for multi-class seizure type classification, considering a dataset with 8 different kinds of epileptic events. They successfully end with a classification accuracy of 88.30% with the Inceptionv3 pretrained network. In [50] a 13-layer CNN is implemented to detect normal, preictal, and seizure classes. The model obtains a final accuracy of 88.67%, with a specificity of 90.00% and a sensitivity of 95.00%. In [51] they proposed a three 1D convolutional layer network to detect the level of consciousness in comatose patients. The problem stated as binary classification between two states (i.e. low consciousness and high consciousness) leads to a final accuracy of 83.3%. A method based on CNN with 2 1D convolutional layers to remove eye blink artifacts is proposed in [52]. The final results outperform those obtained with the well-known independent component analysis (ICA) and regression. In [53] a CNN for binary artifact detection leads to a classification accuracy of 99.20%. A CNN-based models, i.e. DynamicNet, [54], outperformed the current state of the art in motor imagery (MI) classification task. In [55], a CNN is used to classify hand movements from low frequency EEG. Results are compared with two standard ML approaches, i.e. linear discriminant analysis (LDA) and random forest (RF), obtaining comparable or superior results.

Recurrent neural network (RNN)

If NNs try to emulate the brain structure, the recurrent neural networks RNNs try to implement the idea that each sensory state is strongly correlated with the previous ones. The human brain is constantly involved in predictive processing, where past information is used to drive current decisions. In this way, past contextual information can be used during learning. In RNNs, the input is no more

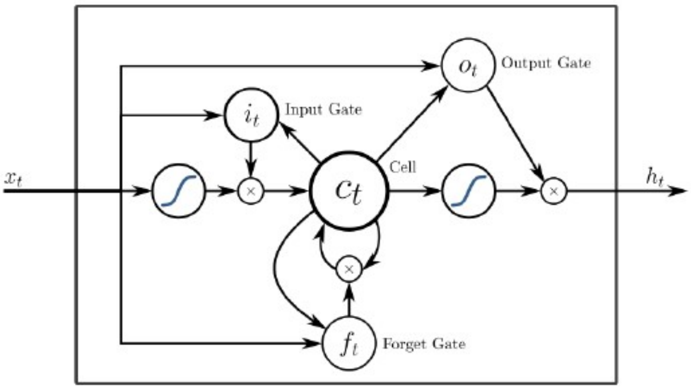


Figure 1.10: The LSTM unit.

static, but it is given by a time series of vectors, and past information is stored in the hidden layer using time-delayed connections. At each time step t, the activation of the hidden layer depends on the current input and on the context layer, which stores the information coming from the previous time step. Then the current hidden unit is used to obtain the output.

$$\begin{cases} h_t = f(W_i x_i + W_c h_{t-1}) \\ o_t = g(W_o h_t) \end{cases}$$

The main drawback of this model is when learning long-term dependencies. As the time sequence starts increasing, the gradient becomes lower and lower resulting in the so called vanishing gradient problem. To overcome this problem, Long Short Term Memory (LSTM) networks [56] have been introduced (Figure 1.10. In this network, a series of gated units learn which and how much information should be retained.

Automatic detection of epileptic events based on LSTM is proposed in [57]. The network, composed of an LSTM layer, a fully connected layer, and a final softmax obtains a detection accuracy of 100%. Moreover, it proves to maintain high detection capabilities also in presence of artifacts. Classification of healthy, ictal, and interictal states is performed in [58]. The use of a two-layer LSTM leads to a final accuracy of 78%. Again. a two-layer LSTM is used in [59]. Classification of three kinds of EEG signal, namely pre-ictal, inter-ictal (seizure-free epileptic), and ictal (epileptic with seizure), leads to an accuracy of 95% while reframing the problem as binary classification (inter-ictal or ictal only detection), the accuracy increases to 98%.

1.3.3 Attention mechanisms in deep learning

The attention mechanism was initially introduced to fix the LSTM troubles in encoding long-term dependencies [6] due to the fixed length of the context vector. The idea is to assign a weight to each encoded hidden state, allowing the decoder to use only the most important one (e.g. that with the highest weight). By identifying some information as more relevant, the fixed length memory space can be optimized. The proposed method is organized into three steps. In the first part, the encoder processes the input and passes to the decoder all the hidden states. Then a score is given to each of them. Each hidden state is multiplied with the correspondent softmaxed score, thus hidden states with a higher score are highlighted. Finally, at each time step, the decoder receives as input a context vector computed as the weighted sum of all the encoded hidden states.

Different kinds of attention mechanisms have been proposed to improve the performances of the existing DL models. In [60] a Convolutional Block Attention Model (CBAM) is proposed. This model aims to emphasize channel and spatial dimension, through the sequential application of a spatial attention module and a channel attention one, to capture *what* is important in the input and *where* it is placed. An LSTM network with attention is proposed in [61] to learn EEG time-series information. Here attention is placed on the top of a 3 LSTM layer network and provides a weight to each LSTM output hidden state. The model, developed for hand movement classification, achieves an accuracy of 83.2%. A transformer model for epilepsy detection is proposed in [62]. The developed model allows detecting the seizure events in 73% of the cases, using only 4 EEG channels. In [63] a Spatiotemporal Attention Network (STAnet) is proposed to decode auditory spatial attention from EEG. The network is composed of three different components: a spatial feature representation, a temporal feature, and a final classification module. Attention is here used to learn which channel and temporal pattern pay attention to. Another novel attention-based architecture (AttnSleep) is proposed in [64] for sleep stage classification. The Temporal Context encoder (TCE) is implemented in the second stage of the proposed architecture and aims to capture the temporal dependencies of the input. In all the tested datasets, they achieved an accuracy higher than 80%. A hierarchical attention block is appended on the top of a CNN to classify epileptic seizure [65]. It is made up of two stages and produces in output a hierarchical attention feature map obtained by multiplying the extracted feature with some attention weights. The outcome achieves an accuracy of 98.33% when discriminating between healthy and ictal subjects, and an accuracy of 95.56% when considering health and interictal classes. In another work, three different attention-based models are tested and compared with two DL approaches on three classification scenarios: neurodegenerative disorders, neurological status, and seizure type [66]. In all three cases, attention-based architectures achieved better results than those obtained with CNN or LSTM only networks.

Chapter 2

Methods

In this chapter, tools, and software employed for this thesis are presented. First, the used datasets are introduced, highlighting the subjects' statistics and the considered classes for the classification. Then the PyEEGLab library [67], the preprocessing pipeline used to process the data and the feature selection process are presented. Finally, the neural network architectures and the metrics employed to evaluate the classification performances are described.

2.1 The datasets

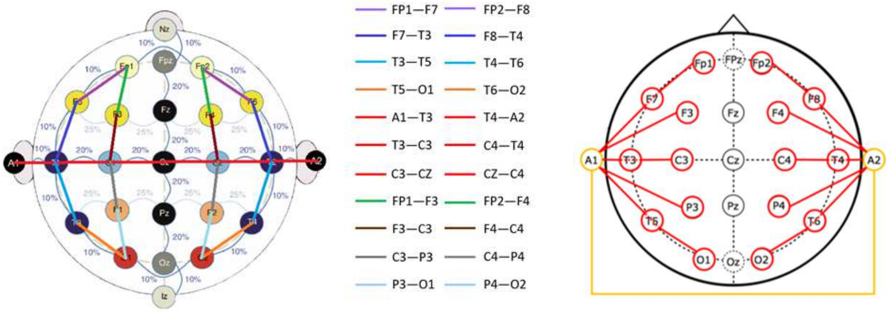
Temple University Hospital EEG (TUH EEG) Corpus is one of the main and largest publicly available databases of clinical EEG data collected from 2002 [8], data that are continuously updated and integrated with new recordings. It aims to provide a huge collection of high quality data for machine learning or deep learning applications, helping researchers solve the lack of data or data generation issues. Moreover, TUH datasets are not only simple EEG collections, but also a neurologist's report including the patient's clinical history, the treatments, the recording description, and impressions is associated with each signal, giving this corpus its uniqueness.
Among all provided datasets three of them are selected, related to three different challenges in EEG research. The first, TUH Abnormal

EEG Corpus (TUAB), contains generic abnormal EEG trials collected from patients, and normal EEG traces from normal subjects [7], the second, TUH Artifact EEG Corpus (TUAR) [8] provides clean EEG signals, and signals corrupted by 5 different kinds of artifacts, the last, TUH Seizure EEG Corpus (TUSZ) [9] consists of EEG trials of patients affected by different types of epileptic seizures.

The vast majority of signals are acquired using the original standard 10/20 system with 21 channels divided into 6 brain regions: prefrontal (Fp), frontal (F), temporal (T), central (C), parietal (P), and occipital (O). Then two reference electrodes (A) are placed on the mastoids. Since data are collected from different hospital units, some recordings could have a higher number of EEG channels, as when high-resolution recording systems are used, or additional electrodes like ECG or EMG for cardiac and muscle activity monitoring. However, not all of those channels will be considered in the following analysis.

Signals are recorded using a bipolar montage to reduce noise, due to the common reference point, and to highlight the events of interest such as epileptic bursts or spiky artifacts [68]. For this purpose a bipolar temporal central parasagittal (TCP) montage is used, computing voltage difference between adjacent electrodes in the longitudinal or transverse direction (known as double-banana). Then, two different kinds of references are considered in the datasets: the average reference (AR), where the reference is given by the average of the electrodes, and the linked ears reference (LE), in which a link between ears is used to create a more stable reference point. However, in this study, just only files acquired with AR montage are used. A visual explanation of the AR and TCP montage is given in Fig. 2.1.

Finally, the used sampling frequency could vary from 250 Hz to 1000 Hz, including intermediate values such as 256 Hz or 400 Hz, but standardization is performed during preprocessing steps, and for all EEG



(a) TCP montage

(b) AR montage

Figure 2.1: AR and TCP montage used for all TUH EEG corpus [68].

signals a 16-bit A/D converter is used.

2.1.1 TUH Abnormal EEG Corpus

The TUH Abnormal dataset (TUAB) [7] contains signals related to normal subjects and several kinds of patients (abnormal samples). Data are collected from a total of 2329 unique subjects. However, 54 of them appear both in normal and abnormal classes, leading to a total of 2383 subjects in the dataset. Moreover, some patients can have multiple sessions, recorded after some months or years, and the total number of recordings is 2993. Differently from the subjects, each session belongs to just only one class. The distribution of subjects and sessions among classes is shown in Table 2.1.

	Abnormal	Abnormal (%)	Normal	Normal (%)	Both	Both (%)	Total
No. subjects	944	40.53	1331	57.15	54	2.32	2329
No. sessions	1472	49.18	1521	50.82	0	0	2993

Table 2.1: TUAB subjects and sessions distribution among classes. White columns contain the number of subjects or sessions in the two classes. Grey columns express the rate of the items in the classes on the whole dataset.

Gender and age distributions are shown in Fig. 2.2 and Fig. 2.3. Sessions are used for this figure since for some patients these can be separated in time by many years, and two recordings of the same subject could belong to two different age groups. Subjects are distributed among all ages from 0 to 100 years, with a higher number of them in the central ages (20-80). Patients are mainly distributed in the range between 40 and 70 years, while normal subjects are in the range of 20-60 years. Regarding gender, the patients' class shows a similar number of female and male subjects, while for the normal class females are higher than males.

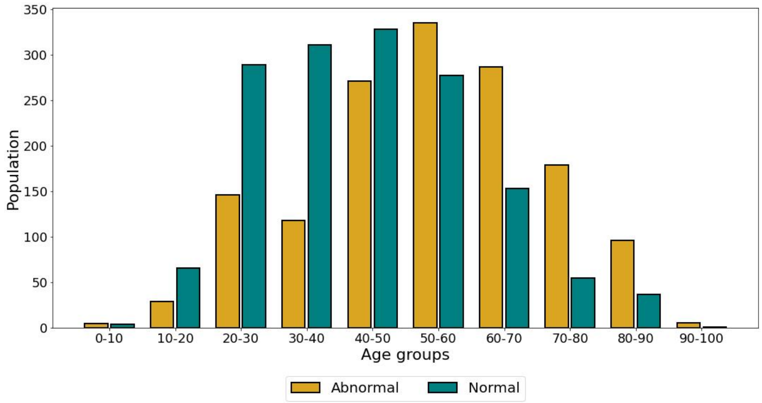


Figure 2.2: TUAB dataset age distribution.

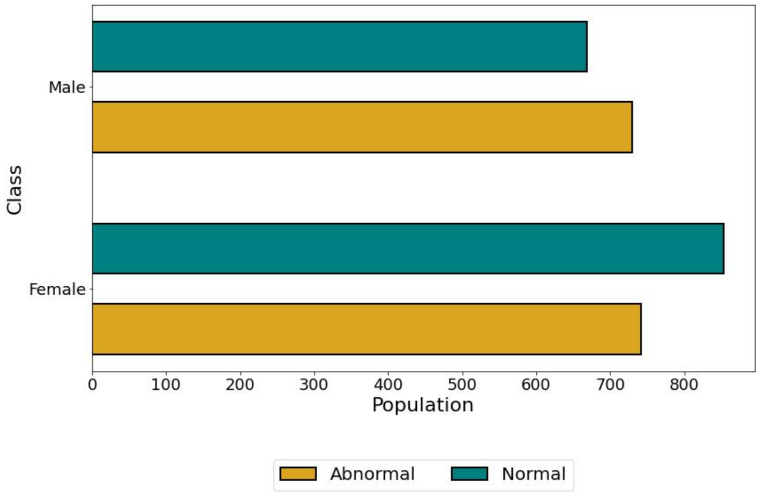


Figure 2.3: TUAB dataset gender distribution.

2.1.2 TUH Artifact EEG Corpus

The TUH Artifact dataset (TUAR) [8] contains both normal and artifactual EEG signals. There are 5 different types of artifacts: chewing artifacts (chew), eye movements (eyem), muscle events (musc), shivering artifacts (shiv), and artifact caused by medical equipment like electrode pop (elpp) or electrode artifact (elec). Moreover, they can occur simultaneously: e.g., eyem-musc, musc-elec, eyem-elec, eyem-chew, chew-musc, chew-elec, eyem-shiv, shiv-elec. Despite the differences among the artifact types, we decided to reduce the classification problem into a binary one: thus we formed an "artifactual" class including all kinds of artifactual samples and a "clean" class with just only clean signals.
The dataset contains EEG signals from 201 unique subjects, 191 of whom appear in both classes. The remaining 10 subjects belong to the artifact class, while there are no subjects only in the clean class. As in the previous dataset, some patients have multiple sessions. In this case, the same session could belong to both classes since part of the signal could be clean and another corrupted by artifacts. The distribution of subjects and sessions among the artifact and clean classes is shown in Table 2.2.

	Artifact	Artifact (%)	Clean	Clean (%)	Both	Both (%)	Total
No. subjects	10	4.97	0	0	191	95.03	201
No. sessions	16	4.73	1	0.39	241	94.88	254

Table 2.2: TUAR subjects and sessions distribution among classes. White columns contain the number of subjects or sessions in the two classes. Grey columns express the rate of the items in the classes on the whole dataset.

Gender and age distribution are shown in Fig. 2.4 and Fig. 2.5 and as for the previous case, they are related to the sessions. Subjects

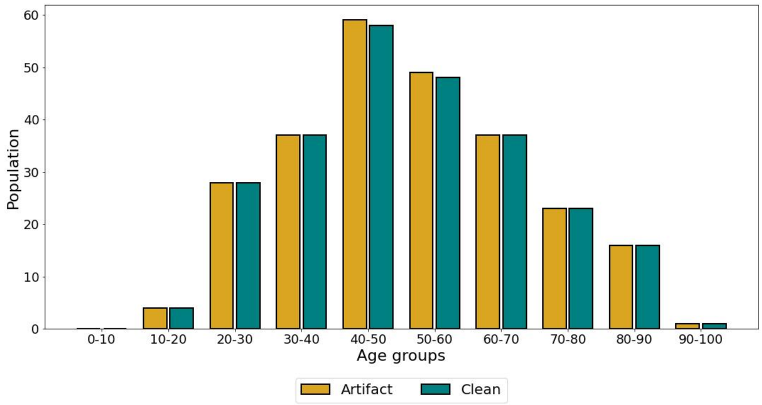


Figure 2.4: TUAR dataset age distribution.

are distributed among all age groups from 10 to 100 years old, in particular in the central classes from 30 to 70 years, and being quite all subjects in both groups, the two distributions are similar. Also for gender distribution, artifacts and clean classes show a very similar number of individuals and the number of females is slightly higher than the males.

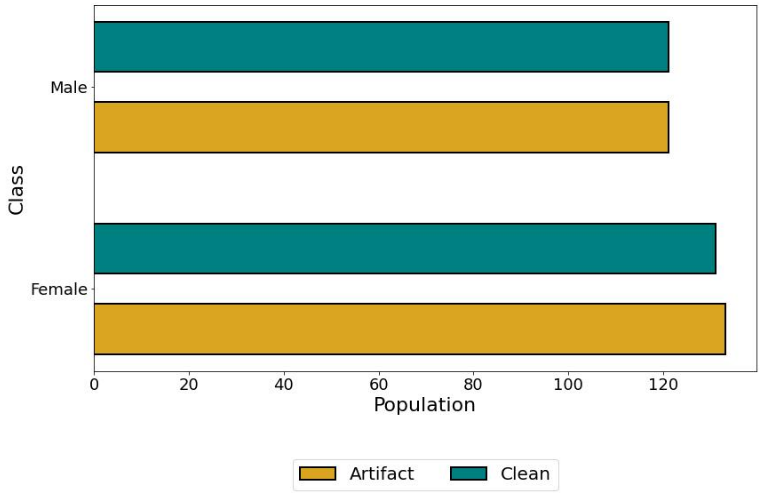


Figure 2.5: TUAB dataset gender distribution.

2.1.3 TUH Seizure EEG Corpus

The TUH Seizure dataset (TUSZ) [9] contains different kinds of events such as absence seizure (absz), clonic seizure (cpsz), focal non-specific events (fnsz), generalized non-specific events (gnsz), simple partial

seizure (spsz), tonic-clonic seizure (tcsz), and tonic seizure (tnsz). However, for this study only focal non-specific and global non-specific seizures are considered as classes for our binary classification.
The EEG recordings come from 153 subjects, 72 of which contribute to the focal class, 38 to the global, and 43 produce samples for both classes, (as reported in 2.3). In some sessions, events from both classes could be recorded. The distribution of subjects and sessions among the focal and global is shown in Table 2.3.

	Focal	Focal (%)	Global	Global(%)	Both	Both (%)	Total
No. subjects	72	47.06	38	24.84	43	28.10	153
No. sessions	161	55.52	86	29.65	43	14.83	290

Table 2.3: TUSZ subjects and sessions distribution among classes. White columns contain the number of subjects or sessions in the two classes. Grey columns express the rate of the items in the classes on the whole dataset.

Unlike the previous two datasets, here there is an unbalance in the number of patients, and sessions with the focal class including almost twice the number of samples of the global class.

Gender and age distribution are shown if Fig. 2.6 and Fig. 2.7. Patients are distributed among all the ages from 0 to 100, predominantly grouped around the central ages (40-70). Regarding ages, both

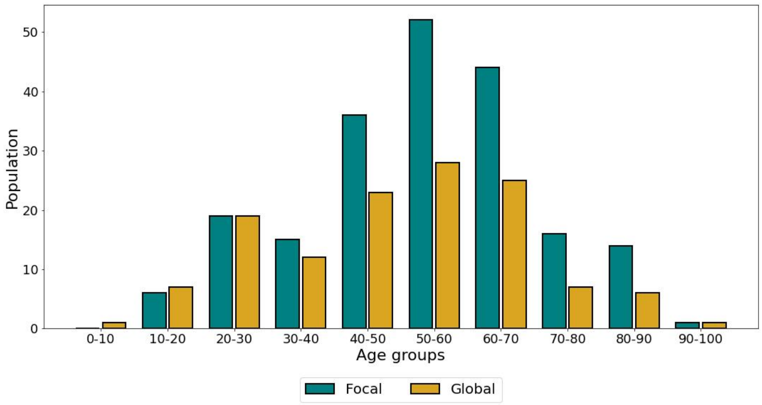


Figure 2.6: TUSZ dataset age distribution.

males and females focal subjects are higher than global ones.

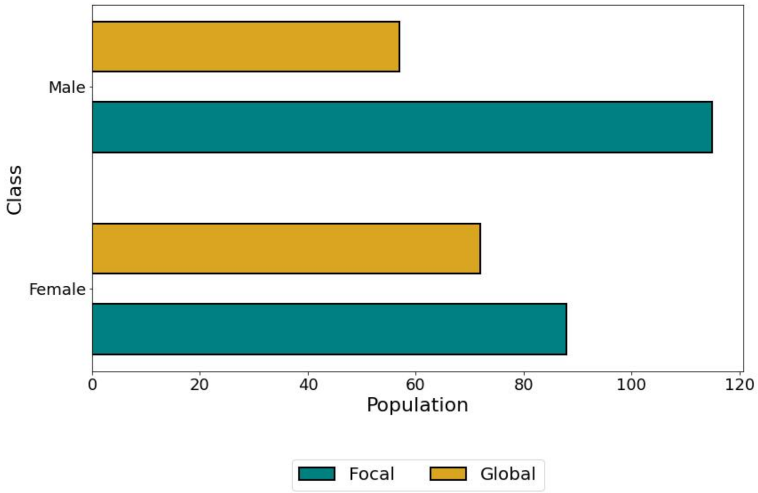


Figure 2.7: TUSZ dataset gender distribution.

2.2 Preprocessing pipeline

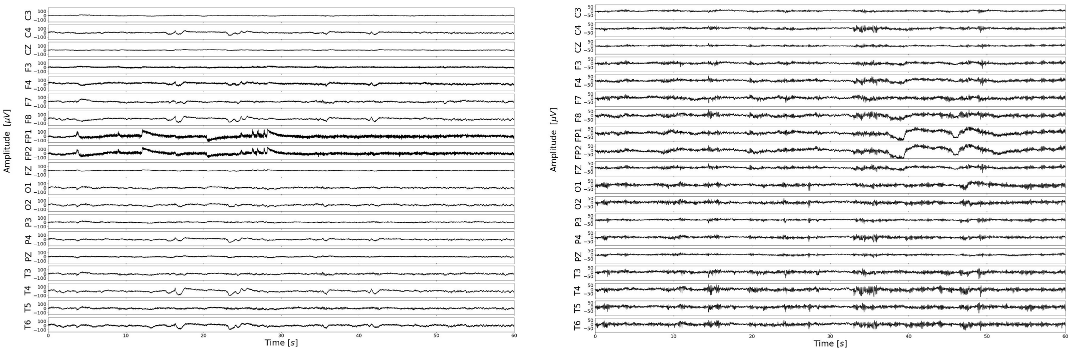
Preprocessing steps and data preparation are performed using the PyEEGLab library, relying on a modified version of the pipeline defined in [69]. Despite the three datasets share the same basic steps, the TUAB pipeline is slightly different from that in TUAR and TUSZ. In the former, each session could belong to just one class, while in the latter each recording can contain multiple chunks of the two classes, and the signal has to be divided into the two parts.

Moreover, some different pipelines that differ just only for the normalization step were tested and described in the next sections. However, only one of them has been used as input to the DL models (see next chapters).

2.2.1 Pre-processing pipeline for the TUAB dataset

Data files are firstly indexed to a cache file and the useful information on the signal characteristics is saved in a SQL database. Such information comprises elements like the list of all channels, the sampling frequency, the label, the minimum and maximum values for each channel, the subject ID, and the specific session number which are required to perform the preprocessing steps.

Then, for each signal, a 60s chunk is cropped between 60s and 120s, in line with [69]. Next, to have as homogeneous data as possible, common channel selection and resampling are performed. In the first step, only channels in common to all subjects are kept and reordered. Also, all signals are resampled to the lowest possible frequency. For this dataset, the common channel set is made of 21 channels: 2 reference electrodes (A1 and A2), 3 central channels (C3, C4, and CZ), 7 frontal channels (F3, F4, F7, F8, FP1, FP2, and FZ), 2 occipital electrodes (O1 and O2), 3 parietal nodes (P3, P4, and PZ), and 4 temporal sites (T3, T4, T5, and T6), while the lowest found sampling frequency is 250 Hz. After that, a bandpass filter with cut-off frequencies of 0.1 Hz and 58 Hz is applied to exclude baseline shift, higher nonrelevant frequencies, and the power line noise. Finally, data are transformed into a Pandas dataframe to make them more easily handled, and z-score within-subject is performed. An example of the 60s signals with the corresponding labels is shown in Fig. 2.8.



(a) Abnormal signal

(b) Normal signal

Figure 2.8: Patient and normal subject 60s signals extracted from the TUAB dataset [7].

The last step of the pipeline is segmentation. Each 60s signal is divided into 30 non-overlapping 2s frames using a sliding window approach. The segment distribution among different subjects and classes is shown in Table 2.4.

subject ID	Abnormal	Abnormal (%)	Normal	Normal (%)	Total
00000016	60	100	0	0	60
00000019	30	100	0	0	30
00000021	0	0	30	100	30
00000039	0	0	30	100	30
00000068	90	100	0	0	90
⋮	⋮	⋮	⋮	⋮	⋮
00000906	60	100	0	0	60
00000929	30	33.33	60	66.67	90
00000930	30	100	0	0	30
⋮	⋮	⋮	⋮	⋮	⋮
00010832	0	0	30	100	30
00010839	0	0	30	100	30
Total	44160	49.18	45630	50.82	89790

Table 2.4: TUAB segment distribution among classes. White columns contain the number of extracted segments for each subject in the two classes. Grey columns express the rate of the segments in a class among all the segments for a subject.

Most of the subjects have frames in just one of the two classes, even if there are some of them with segments in both. There is no big difference between the total number of segments among different subjects, and for each of them there are, in general, 30 60, or 90 segments. Also the total number of segments in the two classes is similar. In total there are 44160 abnormal segments which correspond to 49.18% of the total and 45630 normal ones corresponding to 50.82%.

2.2.2 Pre-processing pipeline for the TUAR and TUSZ datasets

These two datasets differ from the previous one because on the same signal there are chucks belonging to both classes, and a predefined portion of the signal can't be extracted as before. Thus, to the previously stated information, the chunk intervals corresponding to the two classes are stored in the SQL.

Then, the signals are loaded, and again, common channel selection and subsampling are performed. The resampling frequency is 250 Hz for both datasets. The common channel set of TUAR is made up of 23 electrodes: 2 reference electrodes (A1 and A2), 3 central channels (C3, C4, and CZ), 7 frontal channels (F3, F4, F7, F8, FP1, FP2, and FZ), 2 occipital electrodes (O1 and O2), 3 parietal nodes (P3, P4, and PZ), and 6 temporal sites (T1, T2, T3, T4, T5, and T6), while for TUSZ there are 19 channels: 2 reference electrodes (A1 and A2), 3 central channels (C3, C4, and CZ), 6 frontal channels (F3, F4, F7, F8, FP1, and FP2), 2 occipital electrodes (O1 and O2), 2 parietal nodes (P3, and P4), and 4 temporal sites (T3, T4, T5, and T6). Next, filtering is carried out. A wider passband filter with cutoff frequencies of 0.1 Hz and 80 Hz is employed because, by visually inspecting the signals, some interesting components after the 60 Hz are found. A notch filter is then applied at 60 Hz to cut off the power line noise. After that, the signal is cropped and, if more than one interval of the same class is present, the extracted chunks are merged. Finally, data are transformed into a Pandas dataframe, and z-score within-subject is applied. The pipeline ends with segmentation, and each cropped signal is divided into 2s non-overlapping frames. Since signals length are different, from each of them a different number of segments is obtained.

Table 2.5 and Table 2.6 show the segment distribution for classes and subjects for TUAR and TUSZ datasets, respectively.

TUAR dataset has a high variability in the number of segments among subjects: for some of them, the number of artifact segments is very small compared to the clean ones, and for some others is the inverse. Moreover, the total number of extracted segments for each subject span a range between a few hundred and more than one thousand. Also the total number of segments in the two classes is not perfectly balanced. Artifact samples are 40.4% while clean ones are 59.6%.

subject ID	Artifact	Artifact (%)	Clean	Clean (%)	Total
00000254	507	37.81	834	62.19	1341
00000297	4	0.55	720	99.45	724
00000458	45	7.27	574	92.73	619
00000630	274	100	0	0	274
00000647	26	4.33	574	95. 67	600
⋮	⋮	⋮	⋮	⋮	⋮
00005458	103	17,14	498	82.86	601
00005462	1241	99.36	8	0.64	1249
00005649	180	13,12	1192	86.80	1372
⋮	⋮	⋮	⋮	⋮	⋮
00010591	41	5.44	712	94.56	753
00010748	421	56.89	319	43.11	740
Total	66928	40.4	98750	59.6	165678

Table 2.5: TUAR segment distribution among classes. White columns contain the number of extracted segments for each subject in the two classes. Grey columns express the rate of the segments in a class among all the segments for a subject. Red text highlights subjects for which there is a significant unbalance in the number of samples between classes.

subject ID	Focal	Focal (%)	Global	Global (%)	Total
00000016	227	100	0	0	227
00000258	174	32.04	369	67.96	543
00000458	31	40.79	45	59.21	76
00000492	0	0	158	100	158
00001278	14	100	0	0	14
⋮	⋮	⋮	⋮	⋮	⋮
00009232	0	0	51	100	51
00000930	0	0	247	100	247
00009540	0	0	560	100	560
⋮	⋮	⋮	⋮	⋮	⋮
00013095	23	100	0	0	23
00013145	134	100	0	0	134
Total	34000	54.26	28666	45.74	62666

Table 2.6: TUSZ segment distribution among classes. White columns contain the number of extracted segments for each subject in the two classes. Grey columns express the rate of the segments in a class among all the segments for a subject. Red text highlights subjects for which there is a significant unbalance in the number of samples between classes.

Also for the TUSZ dataset, variability in the number of segments appears both among different subjects and in the final number of frames per class. Focal segments are 54.26% of the total while global ones the 45.74%.

2.2.3 On normalization step

Normalization is a common step in EEG preprocessing pipeline [70]. It allows standardizing data among subjects by removing variable, subject dependent, recording effects. Several different normalization techniques could be used like min-max [71], which rescales all the data on the range, the z-score [72] which removes the mean and set the standard deviation to one, or the common average [73] which removes from all the channels the average mean of the signal. In this case, the normalization aim is to reduce the variability introduced by the subjects and to prevent the final classification to be driven by subjects and not by classes.

We try three different normalization techniques, and we report all of them in detail. In the first attempt, min-max centralized normalization is applied. Minimum and maximum values are extracted from each channel of every signal, and then normalization is applied by subject, class, and channels. This means that, for every subject, 21 pairs of minimum and maximum values are extracted (one couple for each channel) as the minimum and the maximum among all the sessions, separately for the two classes, of that subject. Then all these sessions are rescaled with the same pairs of values in the range [-1, 1] according to the formula:

$$data_{norm_{ijk}} = \frac{data_{ijk} - \frac{max_{ijk} + min_{ijk}}{2}}{\frac{max_{ijk} - min_{ijk}}{2}} \tag{2.1}$$

where i is the class, j the subject and k the channel. However, this normalization could lead to some issues. Normalizing the two classes separately could reduce the original differences between them. In fact, in TUAB and TUAR datasets mainly, abnormal and artifact class samples could be characterized by events with higher amplitude (as a spike) that are not found in normal and clean classes. But, when normalizing the two classes separately the amplitude of a spike is rescaled to the same level as the events of the normal trace, which are lower.

The second approach is made up of two steps. First, the common average reference (CAR) is applied [73]: an average of all channels is computed and then subtracted from each channel. Then the z-score normalization considering just only 20% of the subjects. This means that the mean and the standard deviation are computed from a pool of subjects and these values are used to normalize all the other subjects in the dataset according to the following formula:

$$data_{norm_i} = \tfrac{data_i - mean_{pool_i}}{std_{pool_i}} \tag{2.2}$$

where $pool_i$ is the subset of subjects randomly considered for each of the two classes. Also in this case there is an issue related to the CAR step. An abnormal event in one channel (e.g. in frontal or posterior, in case of artifactual signal) could affect the final mean value, which is then used to normalize all the channels (e.g. a central channel). In this way, the event is just only moved between channels but not normalized.

In the last normalization, within-subject z-score is performed. In this approach, the mean and standard deviation are computed for each subject considering all his/her sessions, with no distinction among classes. Then every session of the subject is normalized with mean and standard deviation. In this way, only the variability of the subject is balanced without affecting the variability between classes. Therefore, in the following, we decided to apply this normalization approach.

Table 2.7 summarizes the pipeline steps more compactly.

TUAB	TUAR-TUSZ
1) 60s segment extraction	
2) common channel selection	1) common channel selection
3) resampling	2) resampling
4) filtering (bandpass)	3) filtering (notch + bandpass)
5) within-subject z-score	4) within-subject z-score
6) 2s segmentation	5) single class segment division
	6) 2s segmentation

Table 2.7: Pre-processing pipeline steps for the considered datasets.

2.3 Feature extraction

Once data are processed 11 well-established feature types, both in time and frequency domain, are extracted from each channel of each segment [23]. In the time domain mean, variance, zero-crossings, area under the curve, skewness, kurtosis, and peak-to-peak distance are considered. In the frequency domain, the spectral power using Welch in the four clinically relevant frequency bands, which are delta band (0.5-4) Hz, theta band (4-8) Hz, alpha band (8-12) Hz, and beta band (12-30) Hz are computed.

The final output is a 4D NumPy matrix $A^{n \times m \times f \times c}$ where n is the number of subjects, m the number of segments per subject, f the feature types, and c the channels. Because of the high number of values in the matrix, a visual and statistical inspection of the extracted features is performed to have a preliminary idea of their goodness, and for each feature and each channel separately, the results of the two classes are compared. A visual representation of feature distribution and variability is carried out using the boxplot representation (as it will be reported in Section 4.1), while the two-sided t-test is used to determine whether the two classes' means are significantly different.

2.4 Feature selection

Feature selection is a common step in machine learning. Reducing the number of input variables could be a good solution to reduce the computational cost of the further models as well as to improve their performances, by removing non-informative features [74].

First of all, correlation among all pairs of features is computed. Then a wrapper feature selection method, namely Sequential Feature Selector (SFS) [75] with a Support Vector Machine (SVM) [76] estimator is applied to select the subset of features that leads to the higher classification accuracy. The SVM algorithm is chosen to have a traditional machine learning model with whom to compare the results carried by the new proposed DL architectures.
From the implementation point of view, starting from a 4D matrix is now reshaped into a 2D one to provide a suitable input to compute the correlation and to apply the SVM algorithm. The features set is given by the feature types and the channels, thus each segment of each subject is characterized by a vector:

$$x_s(t) = [f_{11}(t), f_{12}(t), ..., f_{FC}(t)] \tag{2.3}$$

where f_{ij} is the combination of the i-th feature with the j-th channel, with $i = \{1, 2, ..., F\}$ the feature types, and $j = \{1, 2, ..., C\}$ the channels for TUAB, TUAR, and TUSZ dataset respectively, $t = 1, 2, ..., T$ the segment and $s = 1, 2, ..., S$ the subject. The final matrix is then obtained by stacking the features vectors of each trial and subject:

$$M = \begin{bmatrix} f{11}_{11}, & f{12}_{11}, & \cdots & f_{FC_{11}} \\ f{11}_{12}, & f{12}_{12}, & \cdots & f_{FC_{12}} \\ \vdots & \vdots & \cdots & \vdots \\ f{11}_{NM}, & f{12}_{NM}, & \cdots & f_{FC_{NM}} \end{bmatrix} \tag{2.4}$$

where F and C are the features types and channel as in Eq. 2.3,

N the number of subject and M the trials.

In Table 2.8 the sizes of the 2D matrices for the three datasets are given.

	Feature types	Channels	Subject	Segments	Final matrix size
TUAB	11	19	2993	30	89790 × 209
TUAR	11	21	254	variable	165749 × 231
TUSZ	11	17	290	variable	62672 × 187

Table 2.8: 2D input for the correlation and the SVM-SFS algorithm.

2.5 EEG models based on machine learning and deep learning

Four different models are used to perform classification. Two are based on a CNN structure, with and without attention mechanisms, and the others on an LSTM architecture, with and without attention. All the models share a similar architecture that differs in the first layer, where the input information is processed. In CNN and CNN with attention models, a 1D convolution is performed on the input, while in LSTM and LSTM with attention architecture the LSTM layer is used to process the time related information in the input data. The shared structure is made up of an LSTM layer, a dropout, and a dense layer. The classification is performed using the softmax function, which produces a probability distribution for the two output values.

2.5.1 CNN (without attention)

In this first model, data are loaded in batches of size $n \times f$ where n is the number of frames and f the features. Then a lambda layer is used to slice the n frames and each of them is processed with a 1D convolutional layer and then flattened. The n outputs of the flatten layers are all combined and then reshaped. Then the output of this part is passed as input to the structure described before, common to all architectures.

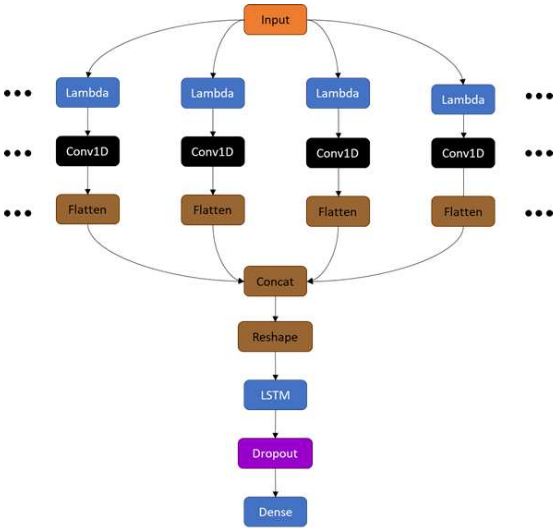


Figure 2.9: CNN architecture.

2.5.2 CNN+Att (with attention)

This model shares the same structure of the previous one with the addition of the attention mechanism, which is placed just after the 1D convolutional layers. The attention layer is based on the Convolutional Block Attention Module (CBAM) [77], designed ad hoc for CNN. It is made up of two attention sub modules: channel and spatial. The former defines the relevant part of the input which is useful, the latter where it is placed. The coefficients matrix is computed using shared Multi Layer Perceptron (MLP) for channel attention, while convolution is used for spatial attention. Then the two sub modules are applied sequentially. The outputs of all the n attention layers are flattened, combined all together, and then passed as input to the structure made up of the LSTM, dropout, dense, and classification layer.

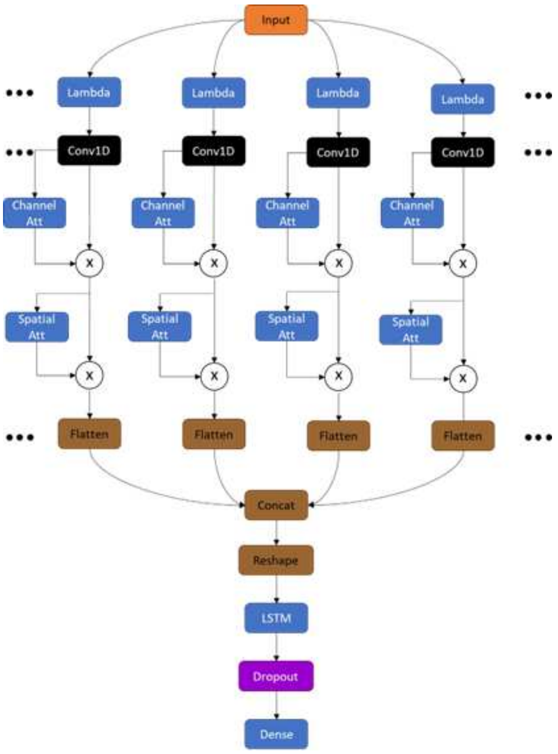


Figure 2.10: CNN+Attention architecture.

2.5.3 LSTM (without attention)

Data are loaded as for the previous models, in a batch of size $n \times f$. A dropout is applied just after the input layer. Then, there are two LSTM layers with dropout. The output of the last layer is fed in input to the dense layer, which performs classification.

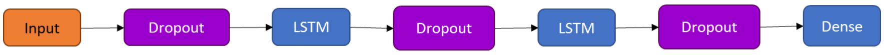


Figure 2.11: LSTM architecture.

2.5.4 LSTM+Att (with attention)

As before, this model shares the same structure of the base LSTM architecture, and the attention layer taken from [78] is placed after the second LSTM layer, just before the dense layer. The LSTM layer builts, for each one of the n input frames, its own representation, then the attention layer assigns a weight to each one of them, giving higher importance to the time step (e.g. the frame) with the most informative content.

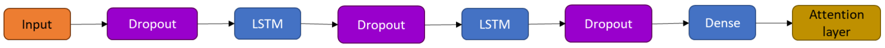


Figure 2.12: LSTM+Attention architecture.

2.6 Training pipeline

The three datasets are split into two parts: training set (90% of the data) and test set (10%). Then, gridsearch is used to optimize the hyperparameters. The considered parameters for the 4 architectures are shown in Table 2.9 for CNN and CNN with attention, and in Table 2.10 for LSTM and LSTM with attention. Each network is trained for 50 epochs, with a batch size of 32, and Adam [79] is used as optimizer.

To increase the reliability of the model, and to obtain a more robust error estimation, a stratified 10-fold cross validation is applied. Moreover, to avoid overfitting early stopping is implemented during training. Validation loss is taken as a reference, interrupting learning after 10 epochs without any improvement in the loss.

Parameters	Values
Learning rate	[0.001, 0.0001]
LSTM hidden units	[32, 64, 128]
Filters	[8, 16]
Kernel size	[3, 5]
Dropout	[0.0, 0.2, 0.4]
Reduction ratio	[8, 16]
Spatial kernel	[5, 7]

Table 2.9: CNN and CNN with attention hyperparameters. Reduction ratio and spatial kernel are optimized just only for the last.

Parameters	Values
Leaning rate	[0.001, 0.0001]
LSTM hidden units	[64, 128]
Dropout input layer	[0.0, 0.2]
Dropout LSTM layer 1	[0.0, 0.2, 0.5]
Dropout LSTM layer 2	[0.0, 0.2]
L2 regularization	[0.01, 0.001]

Table 2.10: LSTM and LSTM with attention hyperparameters.

2.7 Evaluation metrics

Performances are evaluated using the following classification metrics:

- **Accuracy:** it measures the fraction of the corrected classified samples over all the data. It is not very reliable in case of unbalanced data:

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

- **Precision:** it measures the rate of positive samples that are correctly classified and all the samples which are classified as positive. It is the rate of false detection of positive samples:

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

- **Recall:** it measures the ability of the model to correctly identify the true positive samples. It expresses the rate of correct detection of positive samples:

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

- **F1-score:** it is the harmonic mean of the precision and recall. It allows taking into account both type I (false positive) and type II (false negative) errors:

$$F1_{score} = 2 \times \frac{precision \times recall}{precision + recall} = \frac{TP}{TP + \frac{FP+FN}{2}}$$

Chapter 3

Results and Discussion

In this chapter, we present the obtained results. First, we display the statistical analysis carried out on the datasets, then the results of the feature selection procedure. Finally, we summarize the performances obtained with the considered models, and we compare them with those obtained with the SVM algorithm.

3.1 Datasets statistics

Among all possible extracted features, we report in this section the statistical analysis outcomes only for two of them, namely peak-to-peak for the time domain, and delta band spectral power for the frequency domain. We select these two because they show the clearest and most effective results. However, satisfactory results are obtained also for the other features. In the following, we present first the results obtained for the peak-to-peak feature on the three datasets, and then the outcomes for the delta band spectral power feature.

For each dataset, we show the boxplot figures, together with a table holding the number of outliers for each channel and class, and then the p-value. An outlier is a data point that differs significantly from the others. In this thesis, we consider outlier every sample outside the range $\pm k \times IQR$ where $k = 1.5$ and $IQR = Q_3 - Q_1$ the difference between the upper and the lower quantile. The two-sided t-test is performed considering, for each channel separately, the samples of the two classes, and it is used to determine if the means of the two classes are different or not. This is expressed by the p-value and the considered significance level α is set to 0.05.

Fig. 3.1 and Fig. 3.2, together with Table 3.1 show results obtained for the TUAB dataset on the peak-to-peak feature.

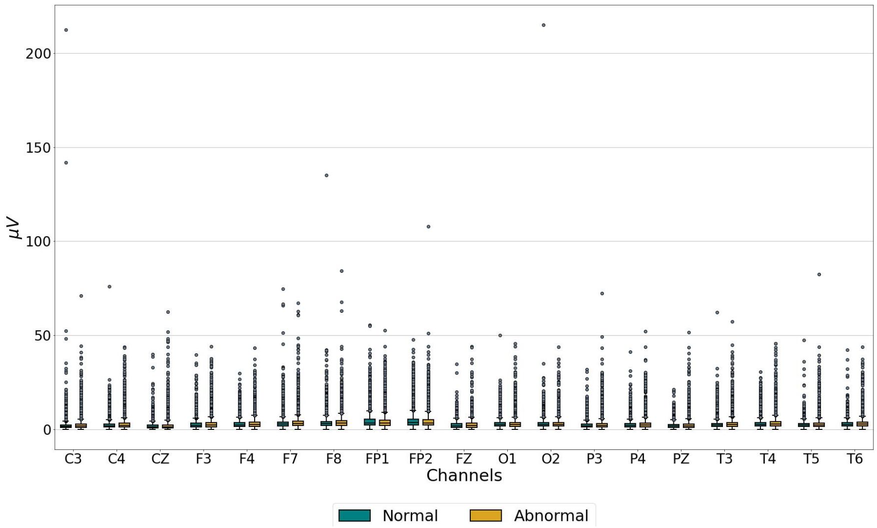


Figure 3.1: TUAB peak-to-peak feature with outliers.

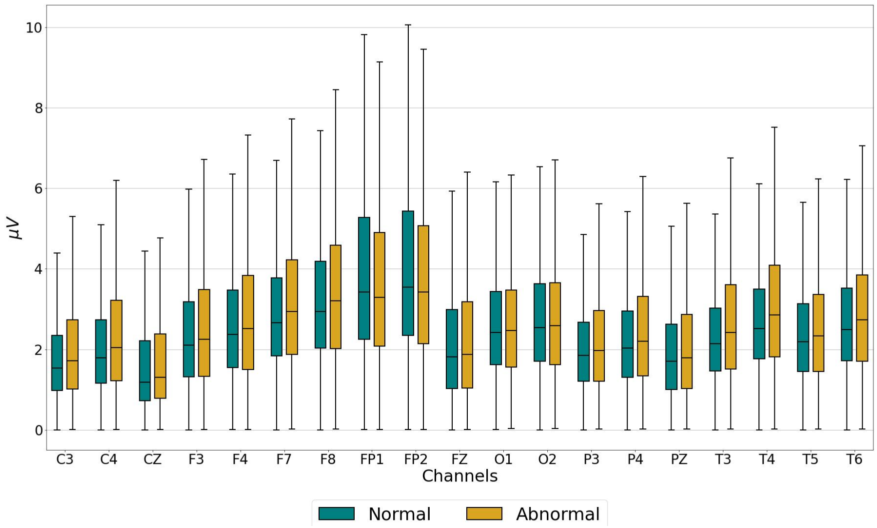


Figure 3.2: TUAB peak-to-peak feature without outliers.

As we can see from the graph and the table, in both classes there are outliers, in a variable percentage from 1% up to 6%, and with more or less the same rate among the two classes. Looking at the distributions of the two classes, they differ except for the O1 and O2 channels, which are characterized by a very similar sample distribution.

Labels	Normal	Normal (%)	Abnormal	Abnormal (%)	p-value
C3	1606	3.52	1636	3.70	< 0.001
C4	1494	3.27	1318	2.98	< 0.001
CZ	2112	4.63	2645	5.99	< 0.001
F3	1097	2.40	1197	2.71	< 0.001
F4	1170	2.56	1079	2.44	< 0.001
F7	1641	3.60	1446	3.27	< 0.001
F8	1546	3.39	1273	2.88	< 0.001
FP1	2828	6.20	2207	5.00	< 0.001
FP2	2702	5.92	2073	4.69	< 0.001
FZ	890	1.95	1089	2.47	< 0.001
O1	1000	2.19	1025	2.32	n.s.
O2	947	2.08	937	2.12	n.s.
P3	938	2.06	1091	2.47	< 0.001
P4	913	2.00	964	2.18	< 0.001
PZ	988	2.17	1197	2.71	< 0.001
T3	1551	3.40	1385	3.14	< 0.001
T4	1184	2.59	1186	2.69	< 0.001
T5	1199	2.63	1161	2.63	< 0.001
T6	959	2.10	951	2.15	< 0.001

Table 3.1: TUAB outliers and p-values with $\alpha = 0.05$. The withe columns display the channels' names, the number of outliers for each class, and the p-value for the t-test. Grey columns show the rate of the outliers on all the samples of a class.

Results for the same feature for the TUAR dataset are shown in Fig. 3.3 and Fig. 3.4, together with Table 3.2. As before, the two class distributions are different, with a wider interquartile range for all channels in the artifact class. In particular, peak-to-peak distances reach higher values for the artifact class.

Outliers span in a range between 4% and 10%, with higher percentage values in artifact class as expected because an artifact could by characterized by a very higher amplitude if compared to a normal sample, and therefore a high peak-to-peak value which can exceed the range of the expected variation. However, an outlier in this class could be not an abnormal value, but just a sample related to an artifact itself, thus characterized by a very high or low amplitude value. Again p-values show that the mean distributions are different.

Finally, the TUSZ dataset results are shown in Fig. 3.5 and Fig.

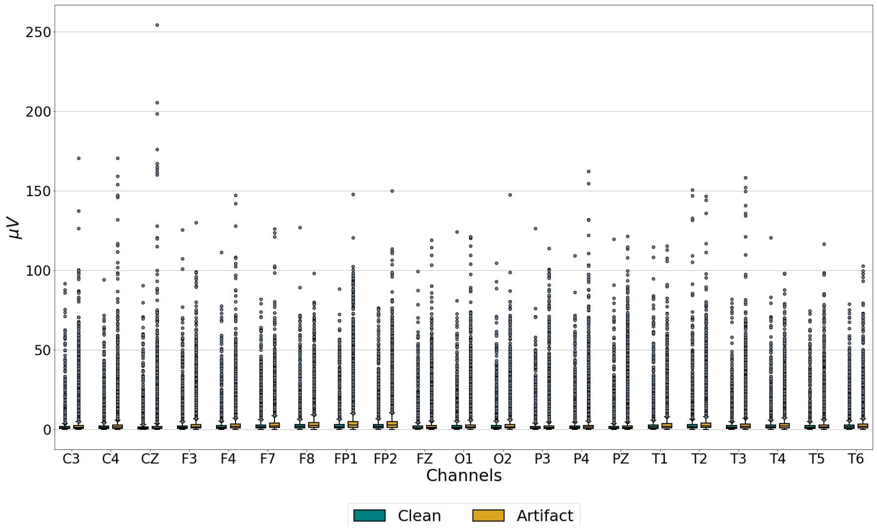


Figure 3.3: TUAR peak-to-peak feature with outliers.

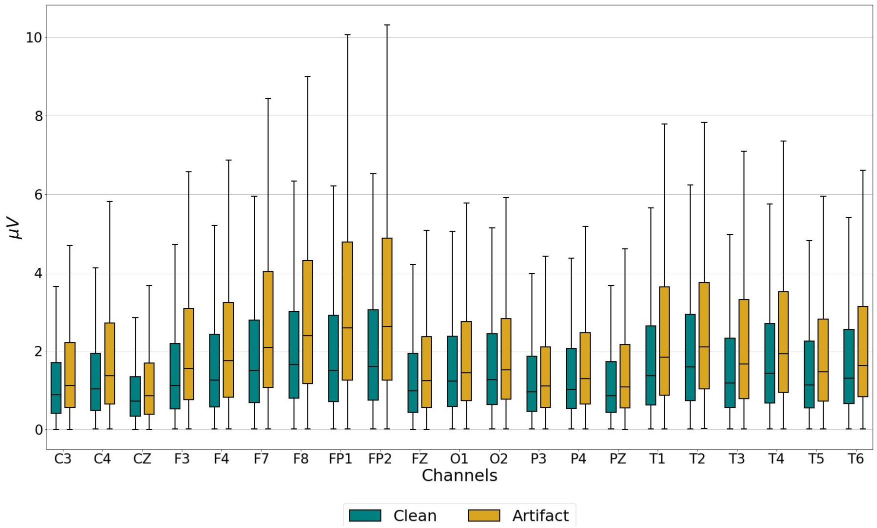


Figure 3.4: TUAR peak-to-peak feature without outliers.

Labels	Clean	Clean (%)	Artifact	Artifact (%)	p-value
C3	3183	3.22	5093	7.61	< 0.001
C4	3267	3.31	4922	7.35	< 0.001
CZ	6934	7.02	4726	7.06	< 0.001
F3	2268	2.30	3216	4.81	< 0.001
F4	1579	1.60	4398	6.57	< 0.001
F7	2382	2.41	4692	7.01	< 0.001
F8	1718	1.80	3782	5.65	< 0.001
FP1	2947	2.98	3981	5.95	< 0.001
FP2	2512	2.54	3655	5.46	< 0.001
FZ	2957	2.99	3620	5.41	< 0.001
O1	3379	3.42	3421	5.11	< 0.001
O2	3535	3.58	3266	4.88	< 0.001
P3	3202	3.24	3892	5.82	< 0.001
P4	3584	3.63	3833	5.73	< 0.001
PZ	4702	4.76	4463	6.67	< 0.001
T1	2686	2.72	4426	6.61	< 0.001
T2	2001	2.03	3811	5.69	< 0.001
T3	3045	3.08	5369	8.02	< 0.001
T4	2090	2.12	4883	7.30	< 0.001
T5	3707	3.75	3788	5.66	< 0.001
T6	2749	2.78	3219	4.81	< 0.001

Table 3.2: TUAR outliers and p-values with $\alpha = 0.05$. The withe columns display the channels' names, the number of outliers for each class, and the p-value for the t-test. Grey columns show the rate of the outliers on all the samples of a class.

3.6, and Table 3.3.

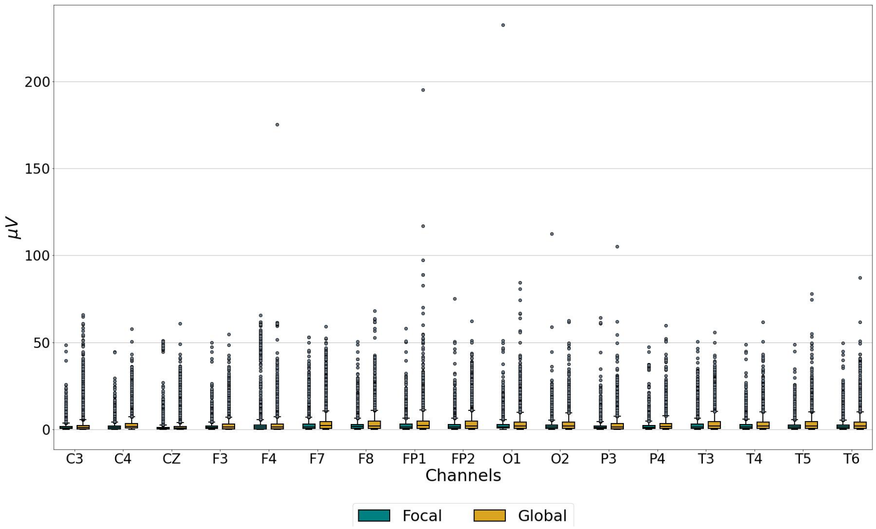


Figure 3.5: TUSZ peak-to-peak feature with outliers.

Outliers are in the range between 1% and 10% with higher value in global class than focal. Boxplot distributions differ as in the two previous cases, meaning that a statistical difference could be found between the two classes. In particular, global samples span a wider range than focal ones. About p-values, they show a value smaller

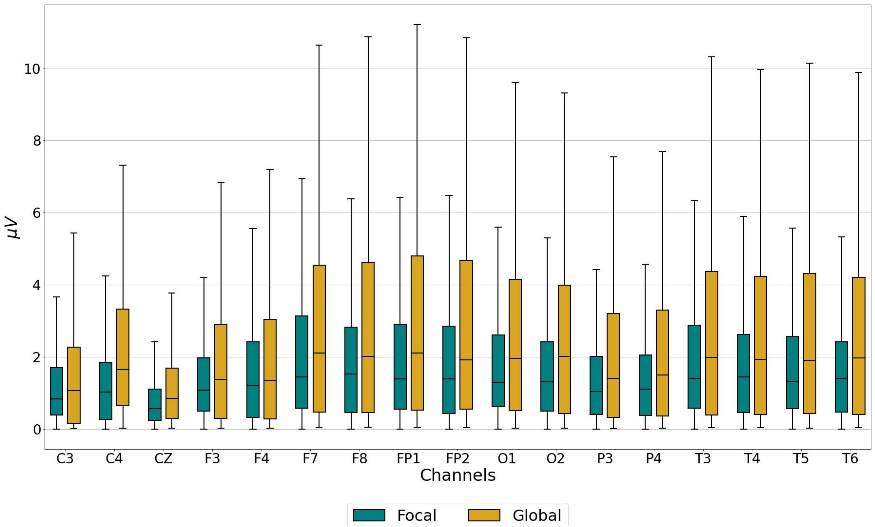


Figure 3.6: TUSZ peak-to-peak feature without outliers.

than 0.001 for all channels and the same considerations done in the previous case can be drawn.

Labels	Focal	Focal (%)	Global	Global (%)	p-values
C3	1519	4.47	2832	9.88	< 0.001
C4	1878	5.52	1487	5.19	< 0.001
CZ	2208	6.49	2762	9.64	< 0.001
F3	2036	5.99	1808	6.31	< 0.001
F4	1192	3.51	2433	8.49	< 0.001
F7	518	1.52	2005	6.99	< 0.001
F8	979	2.88	2430	8.48	< 0.001
FP1	958	2.82	1108	3.87	< 0.001
FP2	1005	2.96	966	3.37	< 0.001
O1	1126	3.31	597	2.08	< 0.001
O2	1493	4.39	553	1.93	< 0.001
P3	1502	4.42	1165	4.06	< 0.001
P4	2312	6.80	926	3.23	< 0.001
T3	581	1.71	1511	5.27	< 0.001
T4	1159	3.41	1856	5.53	< 0.001
T5	989	2.91	730	2.55	< 0.001
T6	1830	5.38	617	2.15	< 0.001

Table 3.3: TUSZ outliers and p-values with $\alpha = 0.05$. The withe columns display the channels' names, the number of outliers for each class, and the p-value for the t-test. Grey columns show the rate of the outliers on all the samples of a class.

Then we show the results obtained for the delta band power spectrum feature. Fig. 3.7 and Fig. 3.8 show results obtained for the TUAB dataset.

Differently from the previous cases, and in general from all time domain features, here boxplots show very few outliers. In particular, no outliers are found in the abnormal class, and few of them (less than 1%) on some normal channels (O2, T5, and T6). The sample

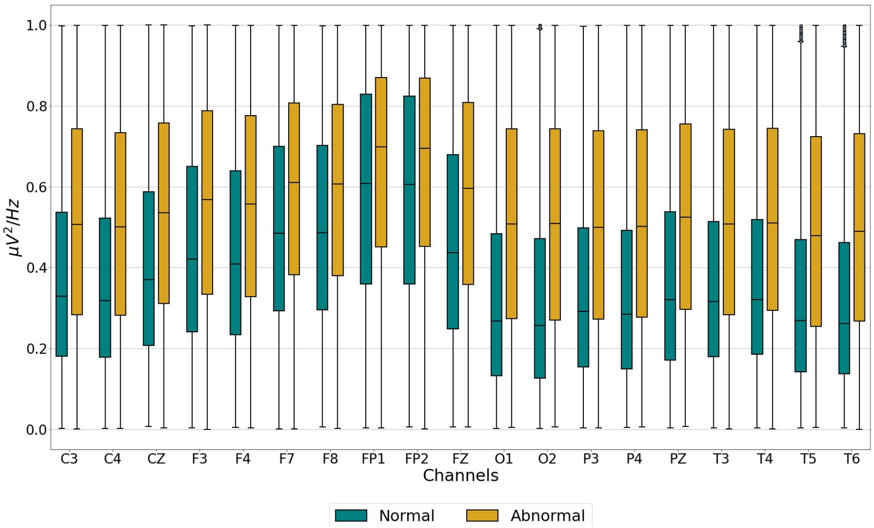


Figure 3.7: TUAB delta feature with outliers.

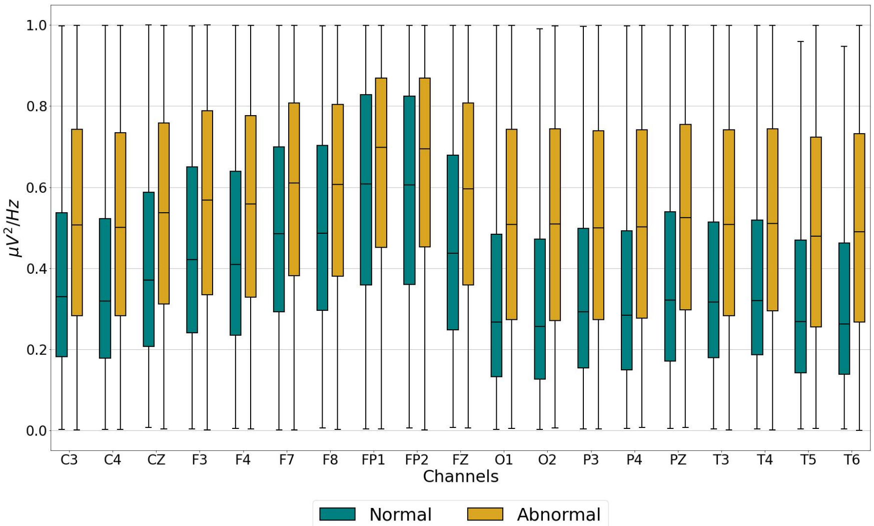


Figure 3.8: TUAB delta feature without outliers.

distributions for the two classes differ in all channels.

Results for the same feature for the TUAR dataset are shown in Fig. 3.9 and Fig. 3.10.

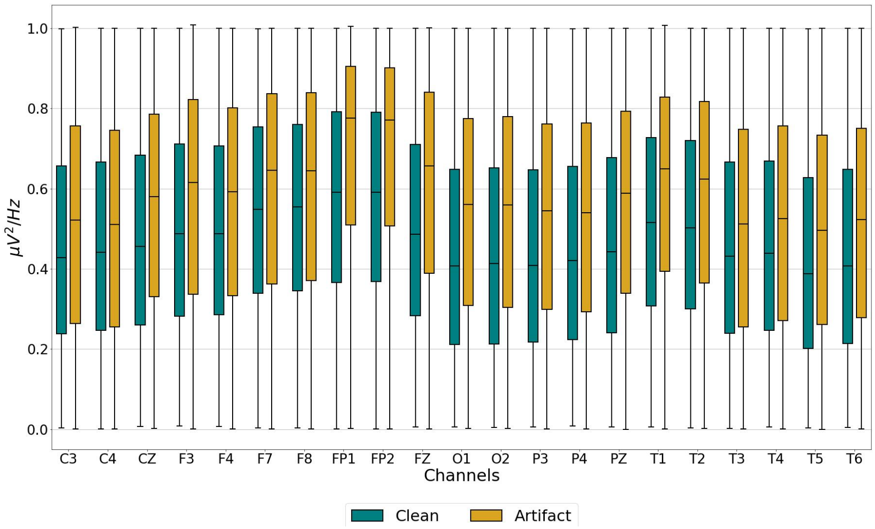


Figure 3.9: TUAR delta feature feature with outliers.

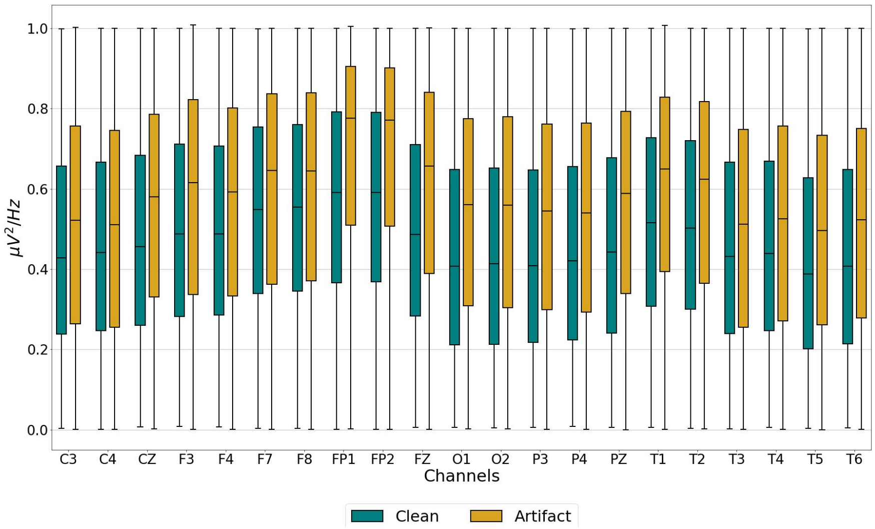


Figure 3.10: TUAR delta feature without outliers.

Also in this case distribution among channels shows some differences, and no outliers are found in either of the two classes.

The last result we report is the TUSZ dataset. Boxplot with and without outliers are reported in Fig. 3.11 and Fig. 3.12 respectively, while outliers number are in Table 3.4.

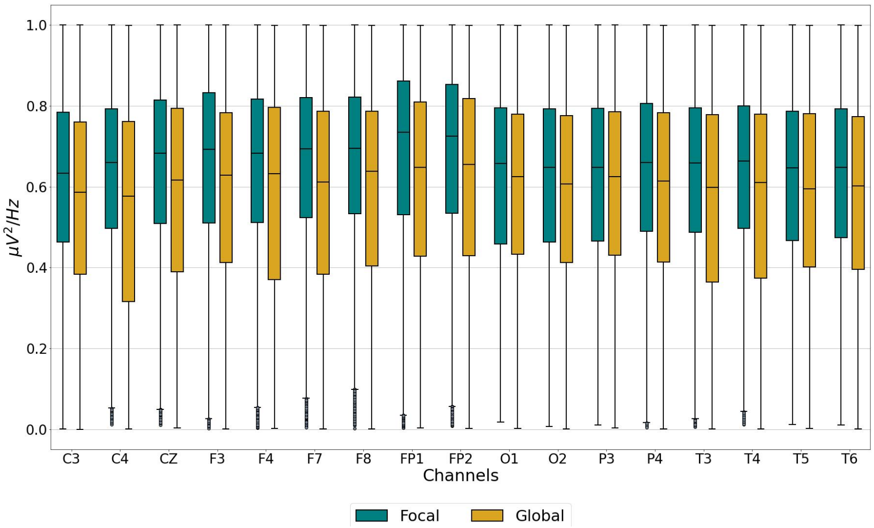


Figure 3.11: TUSZ delta feature with outliers.

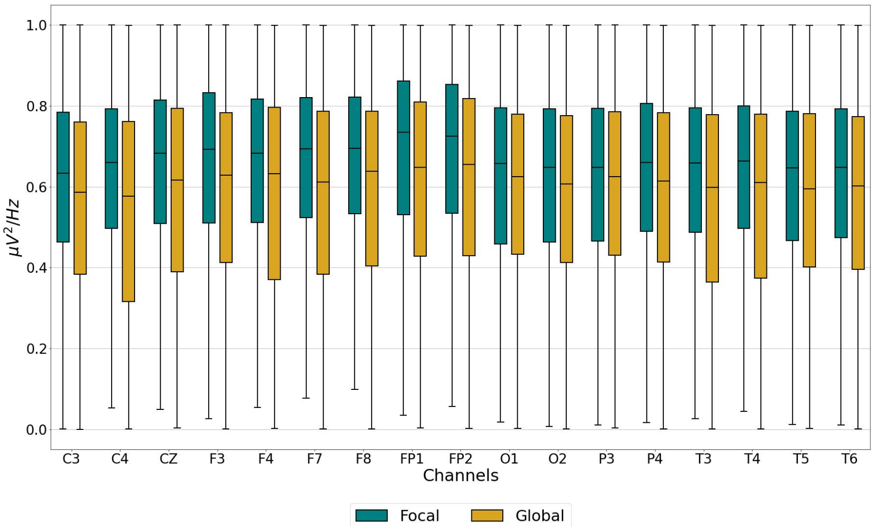


Figure 3.12: TUSZ delta feature without outliers.

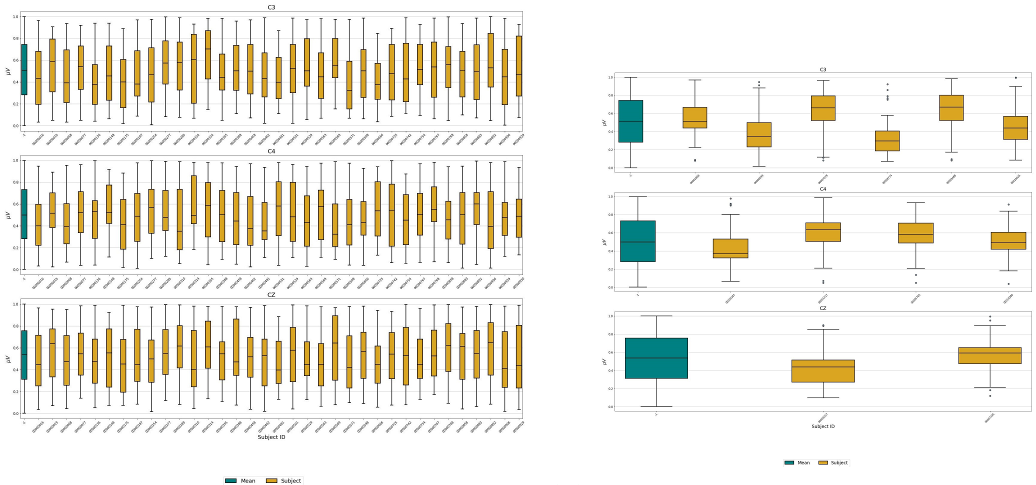
As for the TUAB dataset, the outliers number is small, lower than 1%, and they are only in the focal class. No outliers are found in the global class. Then, as before, distributions differ among classes.

Labels	Focal	Focal (%)
C3	0	0
C4	61	0.18
CZ	105	0.31
F3	65	0.19
F4	206	0.61
F7	304	0.89
F8	200	0.59
FP1	131	0.39
FP2	152	0.45
O1	0	0
O2	0	0
P3	0	0
P4	22	0.06
T3	30	0.09
T4	42	0.12
T5	0	0
T6	0	0

Table 3.4: TUSZ focal class outliers. The withe columns display the channels' names and the number of outliers for the focal class. Grey columns show the rate of the outliers on all the focal class samples. The global class has no outliers in all channels.

Single subject distributions

Then, we want to ensure that the mean distributions previously obtained don't depend on a single subject, or a small group of them. To do this, we select some channels from the previously displayed features, and we plot for each subject a boxplot with only its samples. We keep separated individuals belonging to the two classes and individuals with no outliers and those with at least 5% of outliers. Fig. 3.13 shows the comparison between the mean distribution and the samples' distributions of some randomly selected subjects in the abnormal class for the channels C3, C4, and CZ of the delta band power spectrum feature. Figure 3.13a refers to results for subjects with no outliers while figure 3.13b shows outcomes for subjects with at least 5% od outliers. As we can see, there are no big differences between the mean boxplot (in green) and the single subject ones (in yellow), thus each individual contributes in a similar way to the final output, whether there are outliers or not.



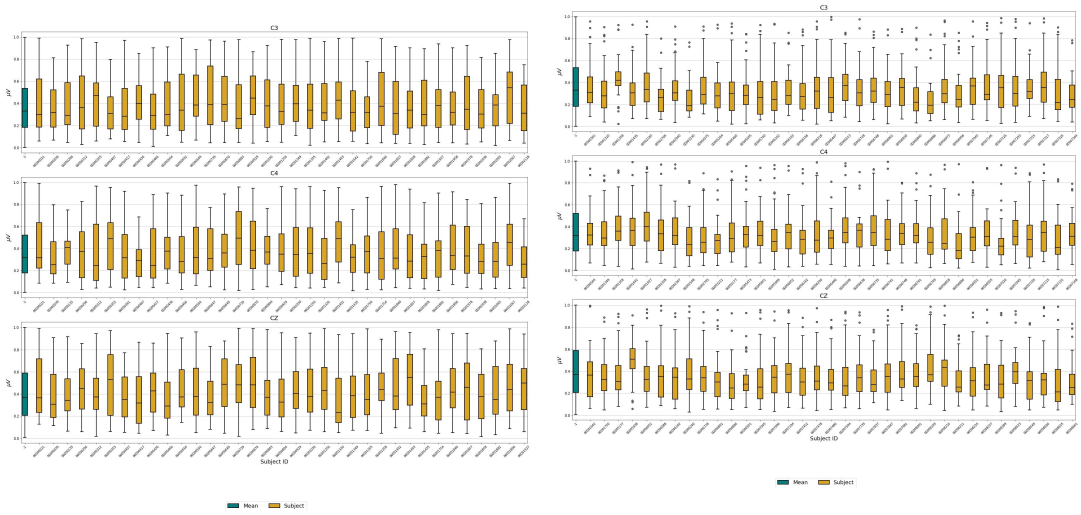
(a) Comparison between mean distribution (in green) and single subject distributions (in yellow) for random subjects with no outliers.

(b) Comparison between mean distribution (in green) and single subject distributions (in yellow) for random subjects with 5% of outliers.

Figure 3.13: Mean distribution and single subject distributions comparison for the abnormal class.

Results for the normal class are shown in Fig. 3.14, and the considerations done for the previous class are still valid.

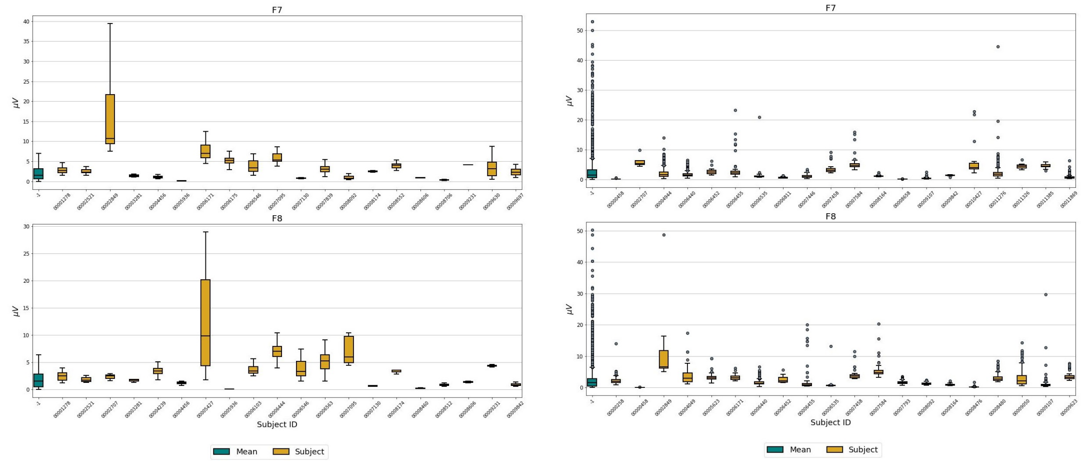
Fig.3.15 and Fig.3.16 show the results obtained for focal and global subjects when considering the peak-to-peak feature and the F7 and F8 channels. Differently from the TUAB dataset, the subject distributions are not always similar to the mean one, and in some cases, they could differ a lot. The same considerations could be done for the global class as we can see in Fig. 3.16.



(a) Comparison between mean distribution (in green) and single subject distributions (in yellow) for random subjects with no outliers.

(b) Comparison between mean distribution (in green) and single subject distributions (in yellow) for random subjects with 5% of outliers.

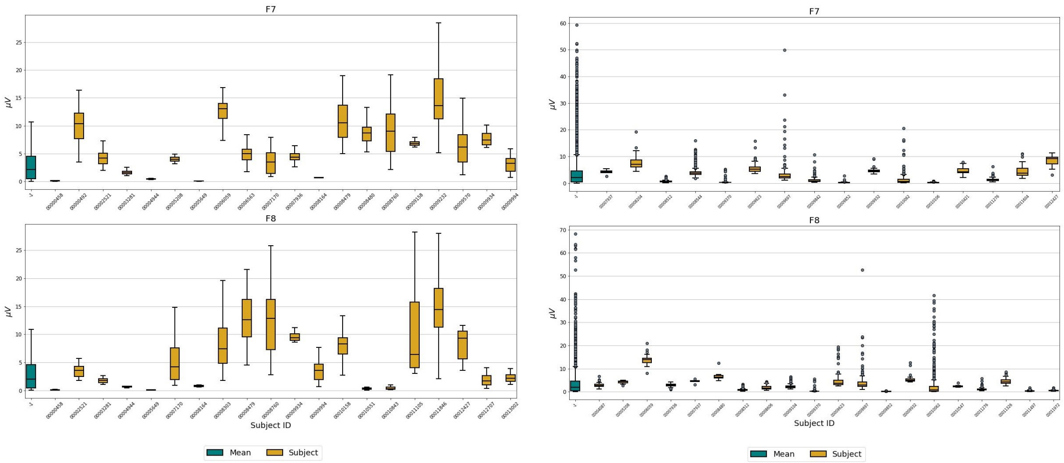
Figure 3.14: Mean distribution and single subject distributions comparison for the normal class.



(a) Comparison between mean distribution (in green) and single subject distributions (in yellow) for random subjects with no outliers.

(b) Comparison between mean distribution (in green) and single subject distributions (in yellow) for random subjects with 5% of outliers.

Figure 3.15: Mean distribution and single subject distributions comparison for the focal class.



(a) Comparison between mean distribution (in green) and single subject distributions (in yellow) for random subjects with no outliers.

(b) Comparison between mean distribution (in green) and single subject distributions (in yellow) for random subjects with 5% of outliers.

Figure 3.16: Mean distribution and single subject distributions comparison for the global class.

In general, TUSZ and TUAR datasets show higher variability in the single subject distribution if compared to the mean one, and this is more evident in the time domain features than in the frequency domain ones.

For this analysis, we decide to report these two features because they are the most representative among all the others. Peak-to-peak outcomes are very similar to those obtained with other time domain features, both for the outliers rate among classes and for how the distributions of the two classes vary between each other. Delta band power spectrum feature is selected because it is the only feature with such a low outliers values, still maintaining a difference in distributions between classes

3.2 Feature selection

3.2.1 Feature correlation

Correlation is computed among all pairs of extracted features. Starting from these last, we create vectors that contain the samples of a specific pair feature type-channel of all the subjects. Then we correlate all these vectors to understand if there are some features that carry redundant information.
Correlation results are reported through a 2D grid heatmap of pixels (as for example Figure 3.17). The two dimensions represent the feature set, while the color is the correlation value. Thus, each pixel of the matrix shows the correlation between a specific couple of features. Because the correlation heatmap is symmetric, we display just only its lower triangular part. In the following, first, we display the whole correlation matrix, then the zoom on some of its portions to highlight the peculiar aspects.

Figure 3.17 shows the correlation results obtained for TUAB dataset.

The correlation matrix is organized as follows:

- pixels on the main diagonals represent the correlation values computed between the same feature.

- triangular submatrices keep the correlation between features characterized by the same feature types and different channels

- elements on the main diagonal of the inner square submatrices refer to correlations computed by features with different types but the same channel

- all other pixels shows the values computed among features with both different types and channels

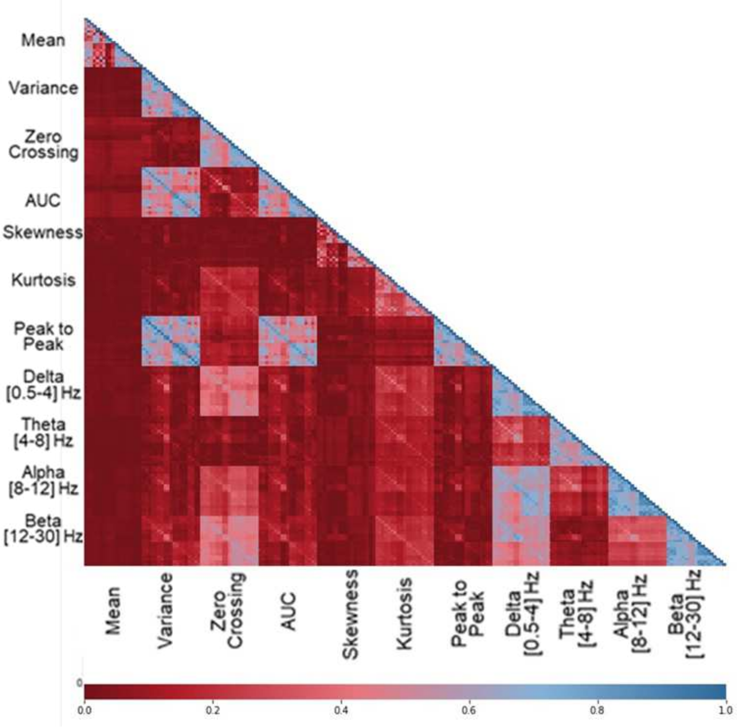
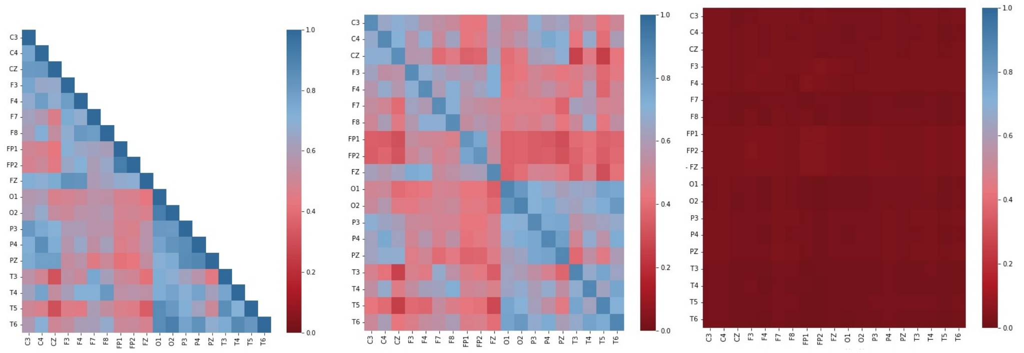


Figure 3.17: TUAB dataset correlation heatmap.

Then, Figure 3.22 shows some zoomed portion of the whole correlation matrix. The first thing we can notice on 3.18a is that the pixels on the main diagonal are blue, corresponding to a correlation value of 1. This is true because values on that line refer to the correlation computed between the same feature. Then, all the other correlation values, are in general higher than 0.4. The same happens in every triangular matrix of the diagonal, thus features that share the same feature type are characterized by higher values. A similar trend could be observed on the main diagonal of the inner square submatrices, which correspond to correlation values between features with the same channel. Regarding all the other values in the matrix, there are some features that are more correlated, such as AUC with variance (Figure 3.18b), peak to peak with variance, and AUC with peak to peak. On the other side, there are many features that have very low correlation values as the mean with all other features (Figure 3.18c), but also skewness and peak-to-peak (except with variance



(a) Variance - Variance correlation

(b) AUC - Variance correlation

(c) Mean - Beta [12-30] Hz correlation

Figure 3.18: TUAB matrix correlation heatmap.

and AUC). Finally, the correlations among frequency domain features show higher values than those between time-frequency pairs.

Figure 3.19 shows results for the TUAR dataset. As in the previous case, the correlations on the main diagonal of the matrix have all values to 1. Triangular submatrices show an even higher correlation if compared with the same elements in the TUAB dataset. As we can see from Figure 3.20a, correlation values are higher than 0.6 in quite all the matrices, and this is true for most of the other matrices on the main diagonal. Correlation between AUC and peak-to-peak (Figure 3.20b), AUC and variance, peak-to-peak and variance show again high values in quite all the pixels, while correlations with mean and with skewness show the lowest value. Regarding correlations between frequency domain features type, they show higher values than those obtained for frequency domain - time domain pairs (Figure 3.20c), but in general lower than 0.5.

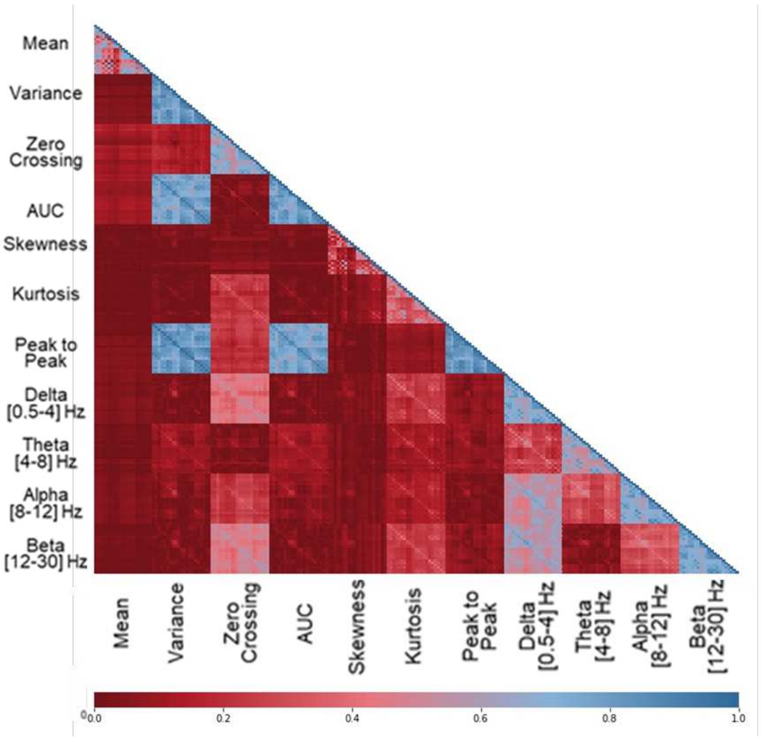
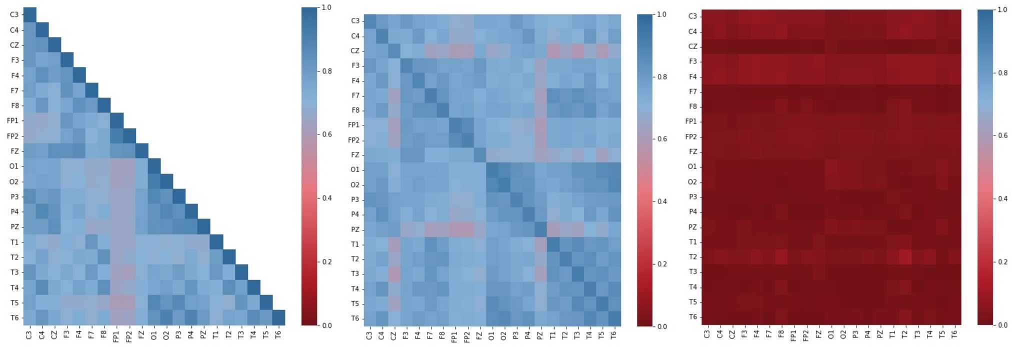


Figure 3.19: TUAR dataset correlation heatmap.

Lately, in Figure 3.21 we found results for the TUSZ dataset.

Similarly to previous results, triangular submatrices show higher correlation values (Figure 3.22c). Squared submatrices between time



(a) Beta [12-30] Hz - Beta [12-30] Hz correlation

(b) AUC - Peak to peak correlation

(c) Skewness - Theta [4 - 8] Hz correlation

Figure 3.20: TUAR matrix correlation heatmap.

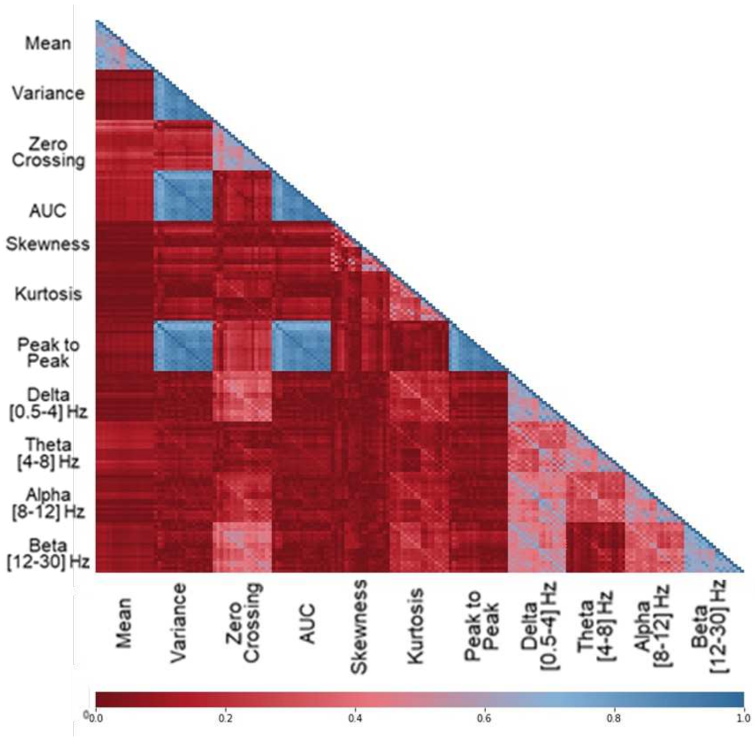
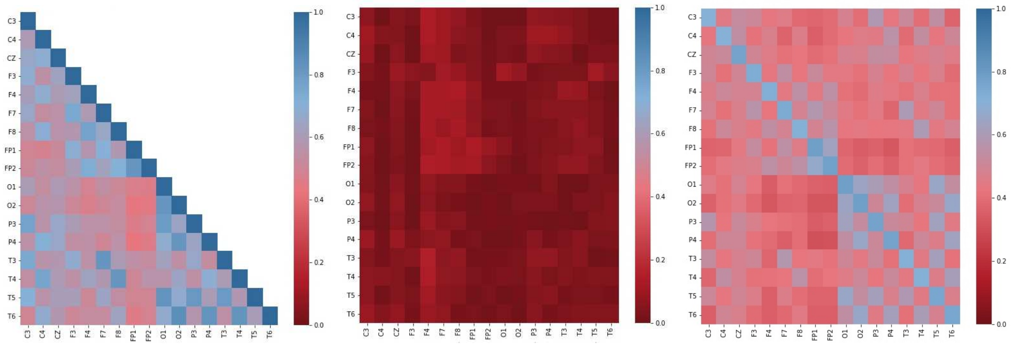


Figure 3.21: TUSZ dataset correlation heatmap.



(a) Delta [0.5-4] Hz - Delta [0.5-4] Hz correlation (b) Beta [12-30] Hz - Skewness correlation (c) Alpha [8-12] Hz - Delta correlation

Figure 3.22: TUAR matrix correlation heatmap.

domain features, or time domain - frequency domain ones have low correlation values (Figure 3.22b), apart from AUC with variance, peak-to-peak with variance, and AUC with peak-to-peak. Again correlations among features in the frequency domain are higher than those in the time domain (Figure 3.22c), excluding the correlation of beta with theta.

3.2.2 SVM-SFS algorithm

The SFS algorithm selects the subset of features that lead to the best accuracy results starting from an empty set and adding one feature at a time. This algorithm uses the SVM as estimator to perform binary classification on the three datasets. The first classification problem refers to abnormalities detection, the second to artifact detection, and the least to seizure type classification.

For the TUAB dataset, we obtain the best results across 24 SVM-SFS iterations, i.e. with 24 selected features (red circle), which corresponds to an accuracy of 73.49% as we can see in Figure 3.23. Then, after adding 5 more features the final accuracy ends to increase significantly thus we stop the algorithm procedure. However, an accuracy of 73.14% is reached already with 15 features (green circle).

The selected features after 15 SFS iterations are: Delta [0.5-4] Hz

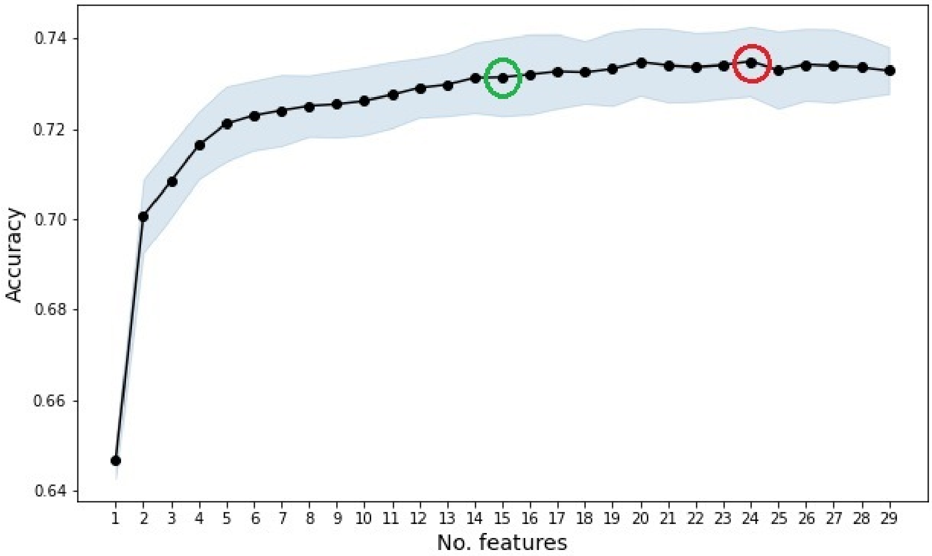


Figure 3.23: TUAB dataset SVM-SFS outcomes. The best classification accuracy is obtained after 24 algorithm iterations.

- O2, Theta [4-8] Hz - O2, Beta [12-30] Hz - P4, Variance - FP1, Variance - F3, Variance - T3, Theta [4-8] Hz - PZ, Variance - C4, Beta [12-30] Hz - O2, Theta [4-8] Hz - FZ, Peak to Peak - FP1, Peak to Peak - F3, Alpha [8-12] Hz - T5, Theta [4-8] Hz - F3, while the configuration which leads to the best performance contains also the following features: Variance - O2, Theta [4-8] Hz - F8, Theta [4-8] Hz - FP1, Alpha [8-12] Hz - C3, Alpha [8-12] Hz - T6, Theta [4-8] Hz - F7, Skewness - F3, Beta [12-30] Hz - FZ, AUC - C4, Peak to Peak - F7.

The results are in line with the literature. In fact, the classification of normal and abnormal EEG could be based on frequencies [80]. The main characteristics of normal EEG in adults include the presence of alpha rhythm, theta activity on frontal and frontocentral region, and a little delta activity [7], [12]. Among the 24 selected features, 14 of them include a frequency band, in particular some alpha rhythms, typical of normal subjects, theta activity, and delta rhythms, that can be found in patients. Moreover, peak-to-peak features, which measure the amplitudes of the signal, are found in the frontal region and can help to discriminate between a normal and abnormal signal, which may include spiky events. Finally, skewness measures the asymmetry in the data, thus the distance from a normal distribution that characterizes the normal traces.

In the TUAR dataset the best final accuracy, that is 75.01%, is reached with 18 selected features (Figure 3.24, red circle.). However, an accuracy of 74.55% is already reached with 12 selected features (green circle). Then, if we increase the features number (up to 5) the final accuracy doesn't increase. Thus, the addition of more information is useless for the classification outcomes.

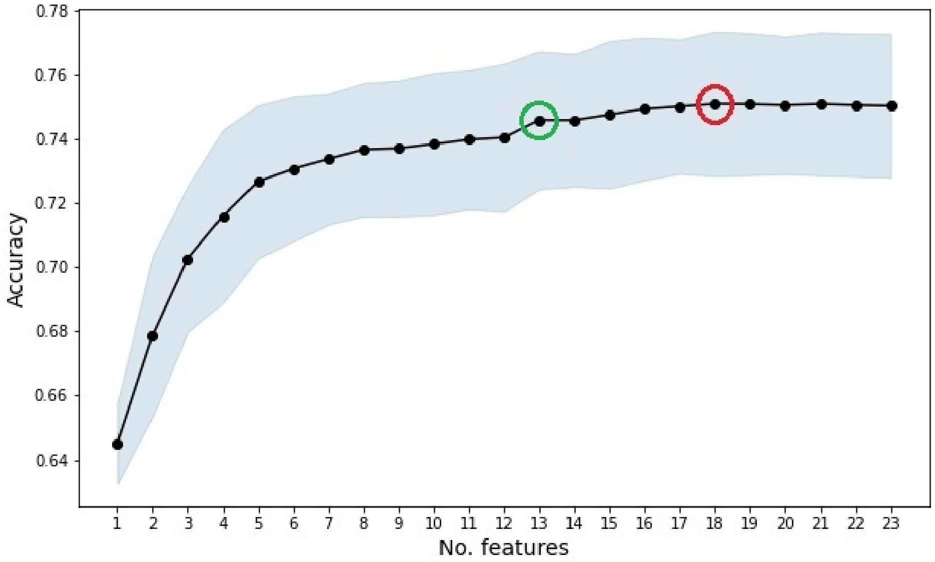


Figure 3.24: TUAR dataset SVM-SFS.

The first 12 selected features are: Peak to Peak - FP2, ZeroCrossing - F8, Delta [0.5-4] Hz - PZ, Skewness - FP1, Delta [0.5-4] Hz - T1, Delta [0.5-4] Hz - FP1, Kurtosis - P4, Beta [12-30] Hz - CZ, Delta [0.5-4] Hz - O1, Peak to Peak - FP1, Beta [12-30] Hz - F4, Peak to Peak - FZ, while the configuration which leads to the best performance contains also the following features: Kurtosis - C4, Beta [12-30] Hz - T6, ZeroCrossing - T3, Variance - F8, Beta [12-30] Hz - T3, Skewness

- F7.

Muscle movements are one of the most common artifacts together with eye movements. The former are characterized by high frequency bursts while the latter show spikes. In addition, both can be found in the frontal region and, only for muscle movement, in the posterior one. Chewing and shivering are similar to muscular artifacts, and they are considered a subclass of them. Instead, electrode pop is characterized by a sudden spike that could generate at any electrode location [81]. In the selected features, we can find 11 of them coming from frontal and frontocentral regions, or posterior ones where muscular and eye movement artifacts are mostly recorded. Then peak-to-peak and zero crossing can capture features characterized by rapid spikes or bursts.

Lastly, Figure 3.25 shows results for the TUSZ dataset.

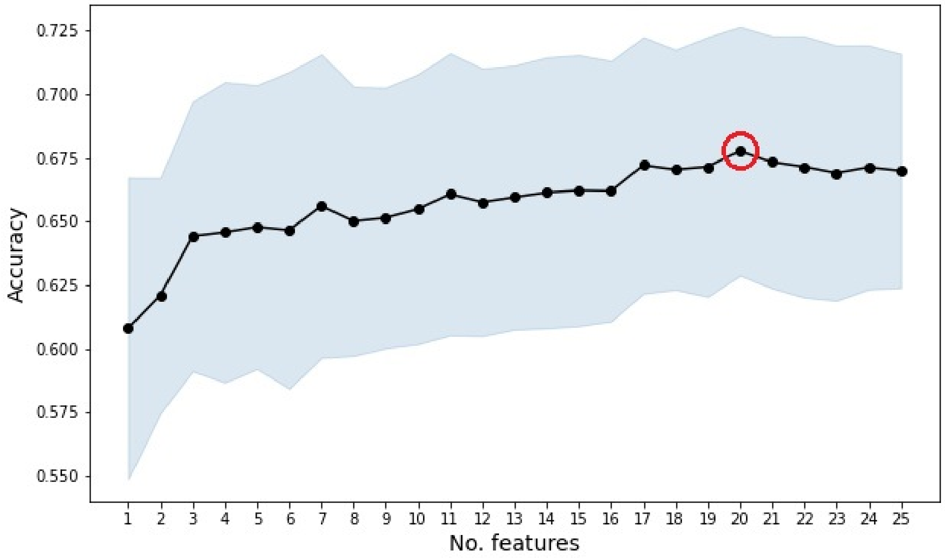


Figure 3.25: TUSZ dataset SVM-SFS.

We found the best accuracy of 67.77% after 20 iterations. However in this case the results are less stable as we can see from the wider standard deviation (the grey area) for each iteration. The selected features are: Peak to Peak - C4, Delta [0.5-4] Hz - C4, Peak to Peak - F4, Peak to Peak - O1, Mean - F4, Delta [0.5-4] Hz - F7, Alpha [8-12]

Hz - P3 Delta [0.5-4] Hz - F3, AUC - F4, Variance - F8, Mean - P4, Delta [0.5-4] Hz - O1, Mean - T3, Kurtosis - P4, Peak to Peak - T5, Skewness - F3, Variance - FP1, Beta [12-30] Hz - P3, AUC - FP2, Mean - F8.

Finally, we run the SVM taking as input all the available features in the datasets to have a traditional ML algorithm as a baseline with which compare the results we will obtain with the DL models. We report the classification metrics, with mean and standard deviation, in Table 3.5.

Model	Accuracy (%)		Recall (%)		Precision (%)		F1-score (%)		No. features
	Mean	Std	Mean	Std	Mean	Std	Mean	Std	
TUAB	72.33	1.55	73.69	9.64	73.19	4.58	72.78	3.57	209
TUAR	72.80	4.69	84.37	6.36	73.87	3.64	78.65	4.00	231
TUSZ	66.25	2.44	71.57	1.95	61.15	1.19	65.98	2.66	187

Table 3.5: SVM classification results (mean and standard deviation reported for each metric).

3.3 Models performances

In the last part of this thesis work, we show the results obtained for the DL models, and we compare them with the SVM ones. For each model, we report the distribution of the values among all the cross-validation runs for the four considered metrics, namely accuracy, precision, recall, and F1-score. Then the final performances of the model are given by averaging all the values among the folds.

3.3.1 CNN (without attention)

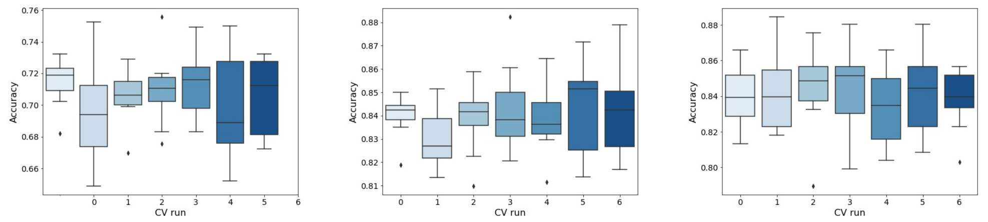
The average results among all the cross validation run for the three datasets are reported in Table 3.6.

Then, the distributions of the accuracy values among all runs are displayed in Figure 3.26

The TUAR and the TUSZ datasets reach good results, with a classification accuracy of 84.36% and 84.96% respectively. On the TUAB

Model	Accuracy (%)		Recall (%)		Precision (%)		F1-score (%)	
	Mean	Std	Mean	Std	Mean	Std	Mean	Std
TUAB	71.41	1.54	82.94	2.53	77.40	2.59	72.45	1.67
TUAR	84.36	1.96	92.19	0.74	87.44	1.29	87.18	1.57
TUSZ	84.96	2.44	89.72	1.95	93.25	1.19	82.98	2.66

Table 3.6: CNN classification results (mean and standard deviation reported for each metric).



(a) TUAB accuracy distribution **(b)** TUAR accuracy distribution **(c)** TUSZ accuracy distribution

Figure 3.26: Accuracy distributions for the cross validation run.

dataset, accuracy stop at 71.41%.

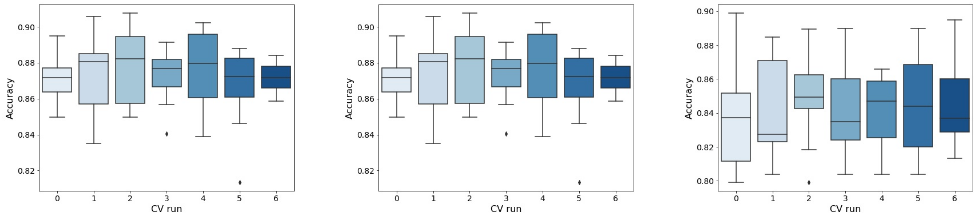
Similar results are obtained in [82]. They compare many different models, such as SVM, kSVM, and 2 different CNN variants (i.e. shallow CNN, and deep CNN) using the TUH corpus dataset. They reached a final accuracy between 65% and 73%. In another work [83] they try to classify 4 different kinds of artifacts using the TUAR dataset. They found an accuracy of 67.59%, obtained by combining classification results coming from three DL models (i.e. CNN, LSTM, and deep CNN). In [84], accuracy values in the range between 62.57% and 71.43% are obtained on the TUAR dataset using different ML algorithms.

3.3.2 CNN+Att (with attention)

In the same way, classification results and the accuracy distributions are reported for the CNN with the addition of specific attention modules (as detailed in Sec. 2.5.2) in Table 3.7 and Figure 3.27.

Model	Accuracy (%)		Recall (%)		Precision (%)		F1-score (%)	
	Mean	Std	Mean	Std	Mean	Std	Mean	Std
TUAB	74.24	2.11	87.02	1.40	83.78	1.59	75.53	2.49
TUAR	87.83	2.11	95.00	0.52	91.84	0.76	89.85	1.81
TUSZ	86.92	2.67	87.18	2.40	89.10	1.52	88.65	3.19

Table 3.7: CNN+Att classification results (mean and standard deviation reported for each metric).



(a) TUAB accuracy distribution **(b)** TUAR accuracy distribution **(c)** TUSZ accuracy distribution

Figure 3.27: Accuracy distributions for the cross validation run.

For this second DL model, the final accuracy increases on all datasets, leading to an average value of 74.24% for the TUAB dataset, 87.83% for the TUAR one, and 86.92% in the TUSZ. Also F1-score values, recall, and precision metrics show an increased value (except for TUSZ precision).

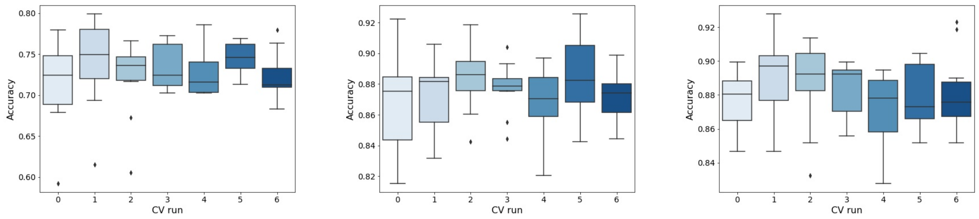
3.3.3 LSTM (without attention)

The LSTM model performances are reported below. Table 3.8 shows the average results among all the cross validation runs for the three datasets, while the accuracy distributions are shown in Figure 3.28.

Model	Accuracy (%)		Recall (%)		Precision (%)		F1-score (%)	
	Mean	Std	Mean	Std	Mean	Std	Mean	Std
TUAB	72.94	3.97	80.14	0.74	68.31	1.30	74.45	4.07
TUAR	87.52	2.34	92.15	0.63	87.35	1.10	90.24	1.62
TUSZ	88.11	2.10	85.17	2.16	91.88	1.51	86.90	2.33

Table 3.8: LSTM classification results (mean and standard deviation reported for each metric).

As in the previous case, the TUAB dataset is the less performing one, with a final accuracy of 72.94%. Moreover, if we look at the stan-



(a) TUAB accuracy distribution **(b)** TUAR accuracy distribution **(c)** TUSZ accuracy distribution

Figure 3.28: Accuracy distributions for the cross validation run.

dard deviation it comes almost to 4% highlighting a high variability during the various runs. This behavior is expressed also by looking at the F1-score. Learning on the TUAR and the TUSZ datasets instead is more stable and lead to an average accuracy of 87.52% and 88.11% respectively.

TUSZ dataset is analyzed in [85]. They used a bi-directional long short-term memory (BiLSTM) reaching a final accuracy of 84.43%. In [86], they achieved accuracy values in the range between 79% and 92%. This last with a novel seizure detection framework, namely channel-embedding spectral-temporal squeeze-and-excitation network (CE-stSENet).

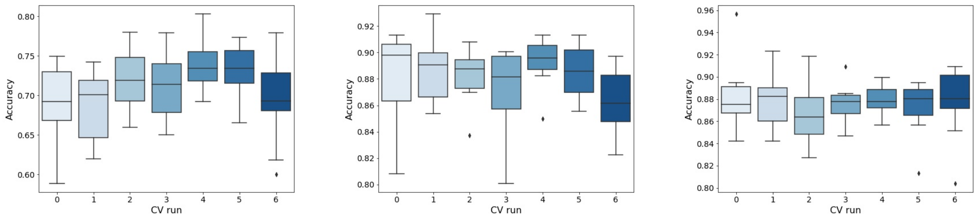
3.3.4 LSTM+Att (without attention)

For this last model, classification results and the accuracy distributions are reported in Table 3.9 and Figure 3.29. Then we compared the results obtained with the simple LSTM (without attention).

Model	Accuracy (%)		Recall (%)		Precision (%)		F1-score (%)	
	Mean	Std	Mean	Std	Mean	Std	Mean	Std
TUAB	74.03	3.33	78.84	1.17	68.36	1.67	74.45	2.28
TUAR	89.36	0.19	92.68	0.58	87.89	0.77	91.50	1.39
TUSZ	88.22	3.03	86.56	1.90	90.63	1.01	87.15	3.00

Table 3.9: LSTM+Att classification results (mean and standard deviation reported for each metric).

Also in this case there is a slight improvement when using the LSTM with attention network for the TUAB and TUSZ datasets.



(a) TUAB accuracy distribution **(b)** TUAR accuracy distribution **(c)** TUSZ accuracy distribution

Figure 3.29: Accuracy distributions for the cross validation run.

The accuracies increase up to 74.03% and 89.36% respectively. The TUSZ dataset is the least affected by the attention mechanism improvement and the final accuracy slightly overtakes that obtained in the simple LSTM (without attention). This might be explained considering that focal and global seizures differ just only for location and duration, while for the other two cases (clean vs artifact, and normal vs abnormal) also the signal amplitude differs a lot. Thus, the attention layer doesn't provide the highest weight to the time steps with the most informative content.

Conclusions and future perspectives

This thesis aimed to compare two DL models, namely CNN and LSTM with their counterparts with attention enhancement. We tested them on three different public EEG datasets which are related to three different challenges in EEG research, i.e. abnormalities detection, artifact detection, and seizure type classification.

In the TUH Abnormal dataset (TUAB), we reach an accuracy up to 74% for the two models with attention, increasing what we have obtained with the counterparts without attention (71.41% and 72.94% for CNN and LSTM).
In the TUH Artifact dataset, the two models with attention outperformed their counterparts without attention, too. Particularly, CNN with attention achieved a final accuracy of 87.83% (84,36% without attention), while the LSTM with attention reached an accuracy of 89.36% (87.52% without attention). These results outperform the models proposed in [83],[84], which provided an accuracy up to 71%.
In the TUH Seizure dataset (TUSZ), the introduction of attention leads to an enhancement only in the CNN-based network, where final performance increases from 84.96% to 86.92%, while for the LSTM-based one, both accuracy values stop at 88% (88.11% for the simple LSTM and 88.22% for the attention-enhanced one). However, the results are in line with the literature. The same dataset is analyzed in [86], where they achieved accuracy values in the range between 79% and 92%. This last value was obtained with a novel seizure detection framework, namely channel-embedding spectral-temporal

squeeze-and-excitation network (CE-stSENet) that first integrates multi-level spectral and multi-scale temporal analysis, then captures hierarchical multi-domain representations with a mechanism based on the squeeze-and-excitation block.

In the future perspective, a third kind of model, namely a Graph Neural Network (GNN) will be introduced, with and without attention [87]. This includes the idea, based on biological evidence, that multiple brain regions are involved during a task. In this way, the interaction between brain areas (i.e. electrodes pair) can be used to extract meaningful features. The strength of this interaction, computed as e.g., the Pearson's correlation between the EEG of two nodes, can be mapped into a graph: each node is an electrode, and each edge is the connection between pairs of electrodes, which is added only if the correlation is strong enough. The attention-enhanced model architecture assigns a relevance coefficient to each feature for a node, allowing it to capture both the relevant network topology and the temporal dependence of the EEG signal, discarding useless information.

Bibliography

[1] E. D. Übeyli, "Combined neural network model employing wavelet coefficients for EEG signals classification," *Digital Signal Processing*, vol. 19, no. 2, pp. 297–308, 2009.

[2] L. A. Gemein, R. T. Schirrmeister, P. Chrabąszcz, D. Wilson, J. Boedecker, A. Schulze-Bonhage, F. Hutter, and T. Ball, "Machine-learning-based diagnostics of EEG pathology," *NeuroImage*, vol. 220, p. 117021, 2020.

[3] P. W. McCulloch, Warren S., "A logical calculus of the ideas immanent in nervous activity," *The bulletin of mathematical biophysics*, vol. 5, pp. 115–133, 1943.

[4] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[5] H. G. E. W. R. J. Rumelhart, David E., "Learning representations by back-propagating errors," *Nature*, vol. 323, pp. 533–536, 1986.

[6] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," 2014.

[7] S. López, G. Suarez, D. Jungreis, I. Obeid, and J. Picone, "Automated identification of abnormal adult EEGs," in *2015 IEEE Signal Processing in Medicine and Biology Symposium (SPMB)*, pp. 1–5, 2015.

[8] I. Obeid and J. Picone, “The temple university hospital EEG data corpus,” *Frontiers in Neuroscience*, vol. 10, 2016.

[9] V. Shah, E. von Weltin, S. Lopez, J. R. McHugh, L. Veloso, M. Golmohammadi, I. Obeid, and J. Picone, “The temple university hospital seizure detection corpus,” *Frontiers in Neuroinformatics*, vol. 12, 2018.

[10] Society for Neuroscience, “Brain facts. a primer on the brain and nervous system.” http://www.brainfacts.org/The-Brain-Facts-Book , 2012. Accessed: 2022-11-07.

[11] S. Ackerman, “Major structures and functions of the brain,” in *Discovering the Brain* (S. Ackerman, ed.), National Academies Press (US), 1992.

[12] P. A. Abhang, B. W. Gawali, and S. C. Mehrotra, “Chapter 1 - introduction to emotion, electroencephalography, and speech processing,” in *Introduction to EEG-and Speech-Based Emotion Recognition* (P. A. Abhang, B. W. Gawali, and S. C. Mehrotra, eds.), pp. 1–17, Academic Press, 2016.

[13] A. Huang, X. Zhang, R. Li, and Y. Chi, “Memristor neural network design,” in *Memristor and Memristive Neural Networks* (A. P. James, ed.), ch. 12, Rijeka: IntechOpen, 2017.

[14] G. Kress and S. Mennerick, “Action potential initiation and propagation: upstream influences on neurotransmission.,” *Neuroscience*, vol. 158, pp. 211–222, 2008.

[15] S. Siuly, Y. Li, and Y. Zhang, *Electroencephalogram (EEG) and Its Background*, pp. 3–21. Cham: Springer International Publishing, 2016.

[16] P. Adjamian, "The application of electro- and magneto-encephalography in tinnitus research – methods and interpretations," *Frontiers in Neurology*, vol. 5, 2014.

[17] C. Demanuele, J. Christopher, and E. Sonuga-Barke, "Distinguishing low frequency oscillations within the 1/f spectral behaviour of electromagnetic brain signals," *Behavioral and Brain Functions*, vol. 3, no. 62, 2007.

[18] G. Buzsáki and A. Draguhn, "Neuronal oscillations in cortical networks," *Science*, vol. 304, no. 5679, pp. 1926–1929, 2004.

[19] J. S. Kumar and P. Bhuvaneswari, "Analysis of electroencephalography (EEG) signals and its categorization–a study," *Procedia Engineering*, vol. 38, pp. 2525–2536, 2012. International Conference on mModelling Optimization and Computing.

[20] V. Jurcak, D. Tsuzuki, and I. Dan, "10/20, 10/10, and 10/5 systems revisited: Their validity as relative head-surface-based positioning systems," *NeuroImage*, vol. 34, no. 4, pp. 1600–1611, 2007.

[21] S. Sanei and J. Chambers, "Introduction to EEG.," pp. 1–34, Wiley online library, 2007.

[22] D. J. McFarland, L. M. McCane, S. V. David, and J. R. Wolpaw, "Spatial filter selection for EEG-based communication," *Electroencephalography and Clinical Neurophysiology*, vol. 103, no. 3, pp. 386–394, 1997.

[23] I. Stancin, M. Cifrek, and A. Jovic, "A review of EEG signal features and their application in driver drowsiness detection systems," *Sensors*, vol. 21, no. 11, 2021.

[24] S. Noachtar and J. Rémi, "The role of EEG in epilepsy: a critical review," *Epilepsy behavior*, vol. 15, no. 1.

[25] I. G. Campbell, "EEG recording and analysis for sleep research," *Curr Protocols Neuroscience*, vol. 49, no. 1.

[26] M. Adamou, T. Fullen, and S. L. Jones, "EEG for diagnosis of adult adhd: A systematic review with narrative analysis," *Frontiers in Psychiatry*, vol. 11, 2020.

[27] F. S. de Aguiar Neto and J. L. G. Rosa, "Depression biomarkers using non-invasive EEG: A review," *Neuroscience Biobehavioral Reviews*, vol. 105, pp. 83–93, 2019.

[28] R. Cassani, M. Estarellas, R. San-Martin, F. Fraga, and T. Falk, "Systematic review on resting-state EEG for alzheimer's disease diagnosis and progression assessment," *Disease markers*, vol. 2018, 2018.

[29] S. L. Oh, J. Vicnesh, E. J. Ciaccio, R. Yuvaraj, and U. R. Acharya, "Deep convolutional neural network model for automated diagnosis of schizophrenia using EEG signals," *Applied Sciences*, vol. 9, no. 14, 2019.

[30] S. Machado, F. Araújo, F. Paes, B. Velasques, M. Cunha, H. Budde, L. F. Basile, R. Anghinah, O. Arias-Carrión, M. Cagy, R. Piedade, T. A. de Graaf, and . R. P. Sack, A. T., "EEG-based brain-computer interfaces: an overview of basic concepts and clinical applications in neurorehabilitation.," *Reviews in the neurosciences*, vol. 21, no. 6, pp. 451–468, 2010.

[31] M. Saeidi, W. Karwowski, F. V. Farahani, K. Fiok, a. H. P. A. Taiar, R., and A. Al-Juaid, "Neural decoding of EEG signals with machine learning: A systematic review.," *Brain sciences*, vol. 11, no. 11, 2021.

[32] S. A. Dudani, "The distance-weighted k-nearest-neighbor rule," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. SMC-6, no. 4, pp. 325–327, 1976.

[33] W. A. Chaovalitwongse, Y. Fan, and R. C. Sachdeo, "On the time series k-nearest neighbor classification of abnormal brain activity," *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, vol. 37, no. 6, pp. 1005–1016, 2007.

[34] K. Rezaee, E. Azizi, and J. Haddadnia, "Optimized seizure detection algorithm: A fast approach for onset of epileptic in EEG signals using gt discriminant analysis and k-nn classifier.," *Journal of biomedical physics engineering*, vol. 6, no. 2, pp. 81–94, 2016.

[35] L. Rosasco, E. D. Vito, A. Caponnetto, M. Piana, and A. Verri, "Are Loss Functions All the Same?," *Neural Computation*, vol. 16, pp. 1063–1076, 05 2004.

[36] C. Y. Sai, N. Mokhtar, H. Arof, P. Cumming, and M. Iwahashi, "Automated classification and removal of EEG artifacts with svm and wavelet-ica," *IEEE Journal of Biomedical and Health Informatics*, vol. 22, no. 3, pp. 664–670, 2018.

[37] A. K. Jaiswal and H. Banka, "Epileptic seizure detection in EEG signal with gmodpca and support vector machine.," *Bio-medical materials and engineering*, vol. 28, no. 2, p. 141–157, 2017.

[38] H. U. Amin, W. Mumtaz, A. R. Subhani, M. N. M. Saad, and A. S. Malik, "Classification of EEG signals based on pattern recognition approach," *Frontiers in Computational Neuroscience*, vol. 11, 2017.

[39] A. Onishi and K. Natsume, "Multi-class ERP-based BCI data analysis using a discriminant space self-organizing map," in *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 26–29, 2014.

[40] M. K. I. Molla, S. K. Saha, S. Yasmin, M. R. Islam, and J. Shin, "Trial regeneration with subband signals for motor imagery classification in BCI paradigm," *IEEE Access*, vol. 9, pp. 7632–7642, 2021.

[41] L. F. Nicolas-Alonso and J. Gomez-Gil, "Brain computer interfaces, a review.," *Sensors (Basel, Switzerland)*, vol. 12, no. 2, pp. 1211–79, 2012.

[42] M. Hosseini, A. Hosseini, and K. Ahi, "A review on machine learning for EEG signal processing in bioengineering," *IEEE Reviews in Biomedical Engineering*, vol. 14, pp. 204–218, 2021.

[43] S. Davidson, N. McCallan, K. Y. Ng, P. Biglarbeigi, D. Finlay, B. L. Lan, and J. McLaughlin, "Epileptic seizure classification using combined labels and a genetic algorithm," in *2022 IEEE 21st Mediterranean Electrotechnical Conference (MELECON)*, IEEE, jun 2022.

[44] H. Robbins and S. Monro, "A Stochastic Approximation Method," *The Annals of Mathematical Statistics*, vol. 22, no. 3, pp. 400 – 407, 1951.

[45] K. Hornik, M. Stinchcombe, and H. White, "Multilayer feedforward networks are universal approximators," *Neural Networks*, vol. 2, no. 5, pp. 359–366, 1989.

[46] P. Cheridito, A. Jentzen, A. Riekert, and F. Rossmannek, "A proof of convergence for gradient descent in the training of arti-

ficial neural networks for constant target functions," *Journal of Complexity*, vol. 72, p. 101646, oct 2022.

[47] L. Patel and K. Goyal, "Applications of artificial neural networks in medical science," *Current Clinical Pharmacology*, vol. 2, no. 3, pp. 217–226, 2007.

[48] K. Fukushima, "Neocognitron," *Scholarpedia*, vol. 2, no. 1, p. 1717, 2007. revision #91558.

[49] S. Raghu, N. Sriraam, Y. Temel, S. V. Rao, and P. L. Kubben, "EEG based multi-class seizure type classification using convolutional neural network and transfer learning," *Neural Networks*, vol. 124, pp. 202–212, 2020.

[50] U. Acharya, S. Oh, Y. Hagiwara, and H. Hong Tan, J. andAdeli, "Deep convolutional neural network for the automated detection and diagnosis of seizure using EEG signals," *Computers in Biology and Medicine*, vol. 100, pp. 270–278, 2018.

[51] Çiğdem Gülüzar Altıntop, F. Latifoğlu, A. Karayol Akın, and B. Çetin, "A novel approach for detection of consciousness level in comatose patients from EEG signals with 1-d convolutional neural network," *Biocybernetics and Biomedical Engineering*, vol. 42, no. 1, pp. 16–26, 2022.

[52] M. Jurczak, M. Kołodziej, and A. Majkowski, "Implementation of a convolutional neural network for eye blink artifacts removal from the electroencephalography signal," *Frontiers in Neuroscience*, vol. 16, 2022.

[53] A. Boudaya, S. Chaabene, B. Bouaziz, H. Batatia, H. Zouari, S. b. Jemea, and L. Chaari, "A convolutional neural network for artifacts detection in EEG data," in *Proceedings of International*

Conference on Information Technology and Applications (A. Ullah, S. Anwar, Á. Rocha, and S. Gill, eds.), (Singapore), pp. 3–13, Springer Nature Singapore, 2022.

[54] A. Zancanaro, G. Cisotto, J. Paulo, G. Pires, and U. J. Nunes, "CNN-based approaches for cross-subject classification in motor imagery: From the state-of-the-art to DynamicNet," in *2021 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*, IEEE, oct 2021.

[55] G. Bressan, G. Cisotto, G. R. Müller-Putz, and S. C. Wriessnegger, "Deep learning-based classification of fine hand movements from low frequency EEG," *Future Internet*, vol. 13, no. 5, 2021.

[56] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, pp. 1735–1780, 11 1997.

[57] R. Hussein, H. Palangi, R. K. Ward, and Z. J. Wang, "Optimized deep neural network architecture for robust detection of epileptic seizures using EEG signals," *Clinical Neurophysiology*, vol. 130, no. 1, pp. 25–37, 2019.

[58] P. Nagabushanam, S. Thomas George, and S. Radha, "EEG signal classification using lstm and improved neural network algorithms," *Soft Computing*, vol. 24, p. 9981–10003, 2020.

[59] M. U. Abbasi, A. Rashad, A. Basalamah, and M. Tariq, "Detection of epilepsy seizures in neo-natal EEG using lstm architecture," *IEEE Access*, vol. 7, pp. 179074–179085, 2019.

[60] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," 2018.

[61] G. Zhang, V. Davoodnia, A. Sepas-Moghaddam, Y. Zhang, and A. Etemad, "Classification of hand movements from EEG using

a deep attention-based LSTM network," *IEEE Sensors Journal*, vol. 20, pp. 3113–3122, mar 2020.

[62] P. Busia, A. Cossettini, T. M. Ingolfsson, S. Benatti, A. Burrello, M. Scherer, M. A. Scrugli, P. Meloni, and L. Benini, "EEGformer: Transformer-based epilepsy detection on raw EEG traces for low-channel-count wearable continuous monitoring devices," in *2022 IEEE Biomedical Circuits and Systems Conference (BioCAS)*, pp. 640–644, 2022.

[63] E. Su, S. Cai, L. Xie, H. Li, and T. Schultz, "Stanet: A spatiotemporal attention network for decoding auditory spatial attention from EEG," *IEEE Transactions on Biomedical Engineering*, vol. 69, no. 7, pp. 2233–2242, 2022.

[64] E. Eldele, Z. Chen, C. Liu, M. Wu, C.-K. Kwoh, X. Li, and C. Guan, "An attention-based deep learning approach for sleep stage classification with single-channel EEG," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 29, pp. 809–818, 2021.

[65] M. S. Chirasani, Sateesh Kumar Reddy, "A deep neural network for the classification of epileptic seizures using hierarchical attention mechanism," *Soft computing*, vol. 26, no. 11, pp. 809–818, 2022.

[66] D. Ahmedt-Aristizabal, M. A. Armin, S. Denman, C. Fookes, and L. Petersson, "Attention networks for multi-task signal analysis," in *2020 42nd Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC)*, pp. 184–187, 2020.

[67] A. Zanga, "Pyeeglab: A simple tool for EEG manipulation." https://github.com/AlessioZanga/PyEEGLab, 2019.

[68] "The temple university hospital EEG corpus: Electrode location and channel labels," *Institute for Signal and Information Processing Report*, vol. 1.

[69] G. Cisotto, A. Zanga, J. Chlebus, I. Zoppis, S. Manzoni, and U. Markowska-Kaczmar, "Comparison of attention-based deep learning models for EEG classification," *arXiv*, pp. 1–10, 2020.

[70] C. Michel and D. Brunet, "EEG source imaging: A practical review of the analysis steps," *Frontiers in Neurology*, vol. 10, 04 2019.

[71] O'Reilly home, "Min–max normalization." Accessed Nov. 24, 2022 [Online].

[72] J. Ko, U. Park, D. Kim, and S. W. Kang, "Quantitative electroencephalogram standardization: A sex- and age-differentiated normative database," *Frontiers in Neuroscience*, vol. 15, 2021.

[73] K. A. Ludwig, R. M. Miriani, N. B. Langhals, M. D. Joseph, D. J. Anderson, and D. R. Kipke, "Using a common average reference to improve cortical neuron recordings from microelectrode arrays," *journal of neurophysiology*, vol. 101, no. 3, 2009.

[74] M. Kuhn and K. Johnson, *An Introduction to Feature Selection*, pp. 487–519. New York, NY: Springer New York, 2013.

[75] F. Ferri, P. Pudil, M. Hatef, and J. Kittler, "Comparative study of techniques for large-scale feature selection* *this work was suported by a serc grant gr/e 97549. the first author was also supported by a fpi grant from the spanish mec, pf92 73546684," in *Pattern Recognition in Practice IV* (E. S. GELSEMA and L. S. KANAL, eds.), vol. 16 of *Machine Intelligence and Pattern Recognition*, pp. 403–413, North-Holland, 1994.

[76] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, 1995.

[77] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," *arXiv*, 2018.

[78] G. Zhang, V. Davoodnia, A. Sepas-Moghaddam, Y. Zhang, and A. Etemad, "Classification of hand movements from EEG using a deep attention-based LSTM network," *IEEE Sensors Journal*, vol. 20, pp. 3113–3122, mar 2020.

[79] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014.

[80] J. W. C. Medithe and U. R. Nelakuditi, "Study of normal and abnormal EEG," in *2016 3rd International Conference on Advanced Computing and Communication Systems (ICACCS)*, vol. 01, pp. 1–4, 2016.

[81] A. Hamid, K. Gagliano, S. Rahman, N. Tulin, V. Tchiong, I. Obeid, and J. Picone, "The temple university artifact corpus: An annotated corpus of EEG artifacts," in *2020 IEEE Signal Processing in Medicine and Biology Symposium (SPMB)*, pp. 1–4, 2020.

[82] D. Nahmias, E. Civillico, and K. Kontson, "Deep learning and feature based medication classifications from EEG in a large clinical data set," *Scientific Reports*, vol. 10, 2020.

[83] D. Kim and S. Keene, "Fast automatic artifact annotator for EEG signals using deep learning," in *IEEE Signal Processing in Medicine and Biology Symposium (SPMB)*, pp. 1–5, 2019.

[84] S. Roy, "Machine learning for removing EEG artifacts: Setting the benchmark," *arXiv*, 2019.

[85] J. He, J. Cui, G. Zhang, M. Xue, D. Chu, and Y. Zhao, “Spatial–temporal seizure detection with graph attention network and bi-directional lstm architecture,” *Biomedical Signal Processing and Control*, vol. 78, p. 103908, 2022.

[86] Y. Li, Y. Liu, W.-G. Cui, Y.-Z. Guo, H. Huang, and Z.-Y. Hu, “Epileptic seizure detection in EEG signals using a unified temporal-spectral squeeze-and-excitation network,” *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 28, no. 4, pp. 782–794, 2020.

[87] I. Zoppis, A. Zanga, S. Manzoni, G. Cisotto, A. Morreale, F. Stella, and G. Mauri, “An attention-based architecture for EEG classification,” pp. 214–219, 01 2020.