POLITECNICO DI TORINO

Master's Degree in Mechatronics Engineering

Master's Degree Thesis

Implementation of a framework to characterize electroencephalographic correlates of walking

Supervisor
Prof. Danilo Demarchi
Prof. Silvestro Micera
Dr. Simone Romeni
Dr. Valeria De Seta

Candidate
Damiano Laudani

Academic Year 2023-2024

Abstract

This thesis investigates the application of machine learning algorithms for the prediction of human gait kinematics utilizing electroencephalogram (EEG) signals, with an initial focus on healthy subjects to evaluate the potential of EEG data in estimating movement during ambulation. Specifically, the research aims to predict the angular dynamics of the joints by employing both EEG signals in isolation and integrating a limited number of kinematic samples. The accuracy of these predictions is assessed in terms of joint angle precision within a defined tolerance, as well as the model's capacity to differentiate between the stance and swing phases, which are crucial for a comprehensive understanding of the gait cycle. The analysis of EEG data obtained from healthy subjects revealed several significant limitations. The EEG signal is highly sensitive to motion artifacts, mechanical disturbances, and muscular interference, complicating the precise extraction of neural activity specifically associated with motor control during gait. Furthermore, the repetitive and cyclic nature of gait performed at a constant speed suggests that the predictability of movement may stem not solely from EEG data, but rather from the inherent periodicity of the motor action itself. This raises pertinent questions regarding the actual informational content of EEG in the context of constant-speed walking, indicating that EEG may offer greater value in scenarios characterized by speed variations or alterations in gait dynamics. In light of these observations, this thesis posits that a promising direction for future research may involve the expansion of experimental protocols to encompass more complex scenarios, such as walking with variations in speed, initiating and ceasing ambulation, as well as navigating inclines and declines. Such situations may activate EEG signals that are more relevant to the regulation of motor control. Additionally, further experimentation could include individuals utilizing robotic-assisted gait (RAG) technologies and patients with spinal cord injuries, thereby assessing the ability of EEG to capture predictive signals in rehabilitation contexts or while utilizing assistive devices, in which EEG may play a pivotal role in facilitating mobility. The findings of this research serve as a foundational step toward the development of novel applications of machine learning in the prediction of gait based on EEG signals, laying the groundwork for future investigations that incorporate more complex gait scenarios and clinical conditions. Ultimately, this work aims to expand the potential applications of EEG signals within rehabilitative and assistive contexts.

Table of contents

Introduction

Over the years, scientific perspectives on human locomotion have evolved, shifting from debates on whether walking and running are primarily governed by peripheral reflexes (Sherrington, 1910), spinal neural networks (Brown, 1911), or the brain and brainstem. For more than a century, researchers have sought to identify the key mechanisms underlying human gait. Advances in experimental techniques and theoretical models have highlighted the interplay of supraspinal commands (Shik et al., 1966, 1968; Shik and Orlovsky, 1976), spinal oscillators, and peripheral reflexes, demonstrating that all are critical contributors to locomotor control. The advent of Mobile Brain/Body Imaging (MoBI) frameworks has provided groundbreaking insights into cortical involvement during locomotion, including walking initiation and ongoing gait. Contrary to earlier views that minimized the role of the brain, recent studies reveal widespread cortical engagement (Widajewicz et al., 1994; Drew wt al., 2002). Key areas such as the anterior cingulate, posterior parietal, prefrontal, premotor, sensorimotor, supplementary motor, and occipital cortices are now recognized as integral to gait regulation (Della Sala et al., 2002; Nutt et al., 1993). Supraspinal locomotor centers, including the brainstem, cerebellum, and cortex, are hierarchically organized to facilitate the integration of multisensory information for dynamic gait control (Fong et al., 2009; Rossignol et al., 2006). Electrocortical activity has been shown to vary with walking speed, stability, perturbations, and adaptive gait patterns, underscoring the complexity of neural contributions to locomotion (Drew et al., 2002; Widajewicz et al., 1994). Despite these advances, decoding gait patterns from brain signals remains a significant challenge. Non-invasive technologies like electroencephalography (EEG) offer promising avenues for developing personalized brain-controlled rehabilitation devices. Leveraging EEG signals within the MoBI framework allows for simultaneous acquisition of EEG and electromyographic (EMG) data with high temporal resolution, providing novel opportunities to analyze movement. Recent studies utilizing MoBI have identified low-frequency EEG potentials as potential indicators of lower-limb kinematics (Presacco et al., 2011; Presacco et al., 2012), although their application in gait decoding remains in its infancy. Only a handful of studies have employed EEG potentials to decode gait events, typically using simple linear classifiers (Kilicarslan et al., 2013; Jorquera et al., 2013). However, such conventional machine learning methods often fail to fully exploit temporal dynamics inherent in locomotor signals. Research on individuals with neurological impairments has further reinforced the significance of cortical and supraspinal mechanisms in locomotor control (Bussel et al., 1996; Calancie et al., 1994; Dimitrijevic et al., 1998; Ferris et al., 2004; Kawashima et al., 2008). Damage to the premotor or sensorimotor cortices is associated with abnormal gait patterns, highlighting the critical role of these

regions (Della Sala et al., 2002; Nutt et al., 1993). Additionally, deficits in executive function and attention have been linked to altered gait dynamics, emphasizing the importance of cortical contributions to real-world locomotion (Laessoe et al., 2008; Sheridan and Hausdorff, 2007; Yogev-Seligmann et al., 2008). Together, these findings suggest that successful gait in everyday life heavily relies on cortical involvement. One major limitation in studying cortical control of locomotion has been the inability to directly measure cortical activity during whole-body movement. Ethical constraints typically preclude invasive cortical recordings in human participants, presenting a challenge for deeper exploration of these mechanisms. Nonetheless, non-invasive approaches like MoBI offer a promising path to bridging this gap (Gramann et al., 2011), advancing our understanding of how the brain controls locomotion in naturalistic contexts.

The thesis begins with a general overview of the brain's role and the main cortical areas involved in controlling human gait, providing a theoretical foundation for understanding the neural mechanisms underlying locomotion. This is followed by an analysis of Mobile Brain Imaging, with a specific focus on the use of electroencephalography (EEG), which has proven to be a key tool for studying, analyzing, and classifying human movement. In this section, the thesis highlights how artificial intelligence, particularly machine learning, has been employed to leverage EEG signals for predicting and classifying movement, presenting the results achieved through various approaches. The central focus of the thesis is the role of EEG in gait decoding, that is, its ability to use brain signals to analyze and classify locomotor dynamics. To explore this topic, the research initially focused on a simple dataset collected from healthy subjects, which included EEG signals and data from inertial measurement units (IMUs). Using this data, machine learning models were developed and tested for predicting joint kinematics, represented by joint angles, and for classifying movement patterns. Recognizing from scientific literature that damage to the motor and premotor cortices is associated with significant alterations in gait dynamics, and highlighting the critical role of these cortical areas in locomotor control, the study progressed to address clinical needs more effectively by implementing an advanced experimental framework. This one integrates not only EEG but also electromyography (EMG), IMU data, and robotic gait assistance via an exoskeleton. Experiments have been conducted on subjects walking with robotic assistance. This new configuration represents a critical step forward, as it aims to study the interactions between brain signals, muscle activity, and kinematics in contexts closer to clinical practice. The data collected from these experiments will provide a foundation for improving rehabilitative and assistive applications, opening new perspectives for understanding and supporting human locomotor control.

1 An Overview of the Nervous System

The nervous system, comprising the central nervous system (CNS) and the peripheral nervous system (PNS), is a complex network that coordinates motor, sensory, and cognitive functions. The CNS, consisting of the brain and spinal cord, serves as the primary control center, while the PNS connects the CNS to muscles and sensory organs via an extensive network of nerves and ganglia. Within the CNS, the brain plays a central role and is divided into the brainstem, cerebellum, and cerebrum. The brainstem, an extension of the spinal cord, facilitates the transmission of information between lower and higher brain regions. The cerebellum, located in the posterior cranial cavity, is essential for motor coordination and balance. The cerebrum, the most complex part of the CNS, is further divided into the telencephalon and diencephalon. The telencephalon includes the cerebral cortex, basal ganglia, amygdala, and hippocampus, while the diencephalon, which contains the thalamus, connects the cortical areas to the brainstem. The cerebral cortex, the outermost layer of the telencephalon, consists of gray matter formed by neurons and glial cells (Figure 1). It is divided into four main lobes, each with distinct functions: the frontal lobe, involved in motor control, language, planning, and decision-making; the parietal lobe, which integrates sensory information and facilitates spatial perception; the temporal lobe, responsible for auditory processing and memory, as well as visuospatial integration; and the occipital lobe, the center of visual processing. Movement-related processes are specifically controlled by motor areas within the cortex, such as the premotor cortex, which identifies the required muscle groups for action; the supplementary motor cortex, which ensures body stability and bilateral coordination; and the primary motor cortex, which transmits motor commands to muscles through descending pyramidal tracts. The motor system, encompassing all CNS structures involved in generating and transmitting motor commands, is hierarchically organized and highly plastic (Figure 2). Higher cortical areas plan and coordinate complex movements, while subcortical structures such as the cerebellum and basal ganglia refine these signals for precision and fluidity. Key circuits include the corticospinal tract, which delivers motor commands to muscles for voluntary movement; the corticocerebellar circuit, which adjusts movements to maintain smooth coordination; and the corticostriatal circuit, critical for initiating repetitive and automatic movements like walking. Sensorimotor circuits further integrate tactile, visual, and proprioceptive information to adapt movements to environmental conditions and maintain balance. Investigating the neural correlates of walking focuses on understanding the activation and interaction of different brain regions during movement. Walking requires dynamic communication between the motor cortex, cerebellum, and basal ganglia. Advanced technologies such as electroencephalography (EEG) enable real-time

monitoring of brain activity during movement, while methods like functional near-infrared spectroscopy (fNIRS) analyze cortical responses during complex motor tasks due to their portability. Other tools, such as functional magnetic resonance imaging (fMRI) and magnetoencephalography (MEG), provide detailed mappings of brain activation, though they may face limitations due to motion-related constraints. These technologies, often combined with electromyography (EMG) to study cortico-muscular connectivity, offer insights into the relationship between neural activity and movement. Such understanding is instrumental in developing innovative and personalized rehabilitation strategies. For instance, robotic walking assistance can be tailored to the patient's neural signals, recovery progress can be monitored through functional connectivity analysis, and targeted interventions can be designed to enhance brain plasticity. This integrated approach, merging neuroscience with movement science, represents a promising frontier for improving mobility and quality of life in individuals with motor impairments.

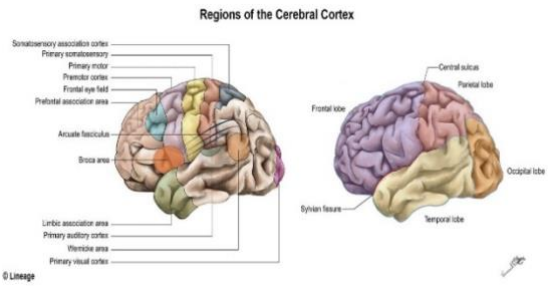

Figure 1:Human cortex lobes

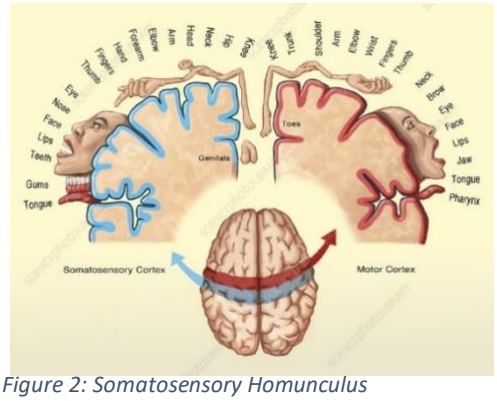

Figure 2: Somatosensory Homunculus

2 EEG and walking

2.1 Electroencephalography

Electroencephalography (EEG) is the only non-invasive neuroimaging technique that utilizes sensors lightweight enough to be worn during ambulation, while offering sufficient temporal resolution to capture intrastride variations in brain activity. EEG records the electrical potentials generated by the brain through electrodes positioned on the scalp. It is a valuable tool for mobile brain imaging due to its non-invasive nature, portability, and direct measurement of cortical

activity, unlike indirect methods such as those based on blood flow dynamics (e.g., fNIRS and fMRI), which reflect neuronal metabolic processes. However, a significant challenge in employing EEG during full-body motion lies in its inherently low signal-to-noise ratio. This limitation arises because the signal must traverse multiple biological layers, including the skull, meninges, cerebrospinal fluid, and skin, as depicted in Figure 3. Consequently, the amplitude of the recorded electrocortical signals is minimal and frequently masked by substantial motion and muscle artifacts generated during activities such as walking.

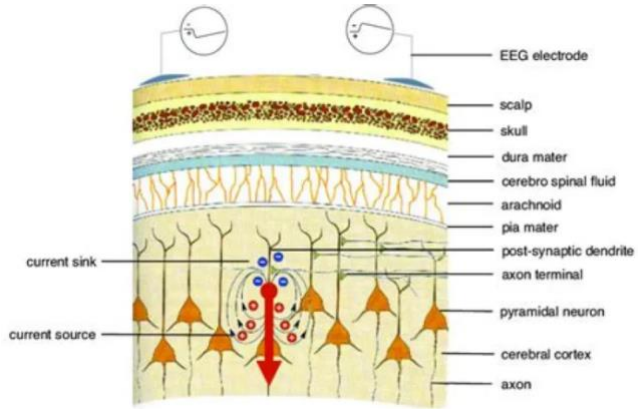

Figure 3: Neural sources EEG

This activity is captured via electrodes placed on the scalp following the international 10-20 system, which ensures standardized electrode placement for consistent and reproducible measurements (Figure 4).

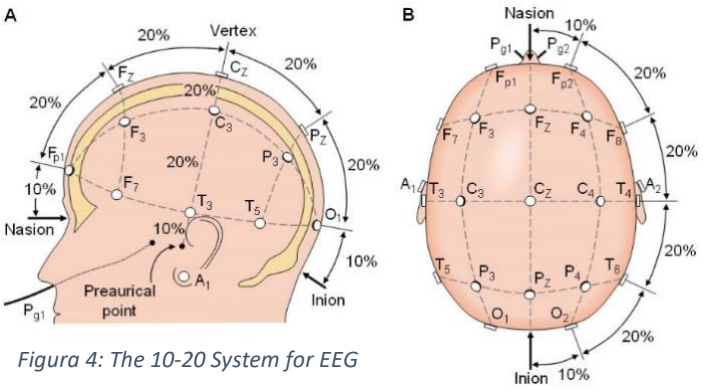

Figura 4: The 10-20 System for EEG

EEG electrodes detect minute electrical charges (10–100 microvolts) generated by the activity of neuronal populations. Neurons communicate via electrical impulses and chemical signals, creating electrical activity that can be measured using EEG. Although the electrical potential of an individual neuron is too small to be detected at the scalp, the synchronized activity of a large group of neurons produces far-field potentials that can propagate to the scalp and be recorded. Scalp EEG primarily captures electrical activity originating from cortical structures located beneath the skull. Neuronal communication occurs through synaptic activity and oscillatory electrical currents, conveying information related to motor commands, sensory input, and cognitive processes. EEG signals primarily arise from post-synaptic currents and action potentials (Thio and Grill, 2023), with cortical pyramidal neurons being the dominant contributors (Teplan, 2002). These neurons are oriented perpendicular to the cortical surface and generate electrical currents directed either toward or away from the scalp (Hari and Puce, 2017). Their consistent alignment enables the summation of their activity, making it detectable via EEG. In contrast, non-pyramidal neurons and cells in deeper brain structures contribute minimally to the recorded signals (Hari and Puce, 2017). Changes in EEG signals are often characterized by synchronization and desynchronization (Pfurtscheller and Lopes da Silva, 1999). Synchronization refers to a state where a neuronal population exhibits increased congruence in the timing of postsynaptic potentials compared to a baseline, or when a greater number of neurons contribute to such congruent timing. Conversely, desynchronization occurs when neurons show less alignment in their postsynaptic potential timing or when fewer neurons are engaged in synchronized activity relative to the baseline state. Desynchronization can result either from more independent neuronal firing patterns or a reduction in the number of neurons participating in congruent activity. The power and frequency of EEG signals have long been used as markers of various brain states, particularly during rest (Teplan, 2002). EEG signals are commonly categorized into five frequency bands, each associated with distinct neural activity: delta (<3.5 Hz), theta (4–7.5 Hz), alpha (8–13 Hz), beta (14–35 Hz), and gamma (>35 Hz) (Hari and Puce, 2017) (Figure 5).

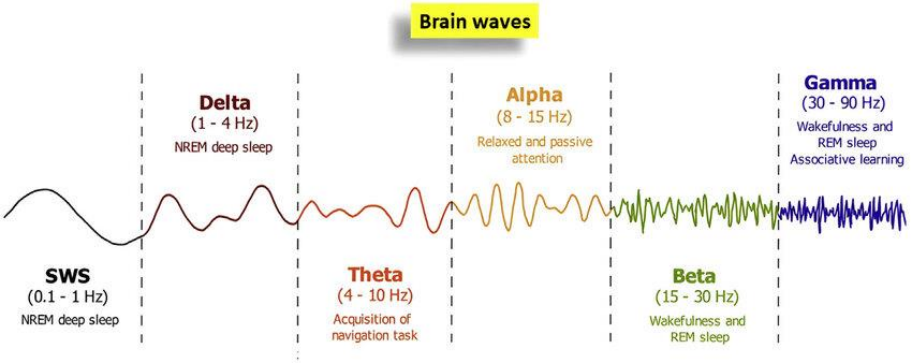

Figura 5: EEG frequency bands

Different frequency bands are thought to be correlated with certain behavioral mental states in humans participants. Brain activity can also be analyzed as event-related potentials that are time-locked to a stimulus (Luck,2014). An evoked potential is a fluctuation in voltage that was caused by an external or internal stimulus (Bickford, 1987; Nunez and Srinivasan, 2006). Event-related potentials are extracted by averaging epochs of EEG that are time-locked to an event (Gevins and Remond, 1987; Teplan, 2002). Any spontaneous fluctuations unrelated to the event are averaged out, leaving only the activity which is consistently associated with the processing of stimulus (Teplan, 2002). Event-related spectral perturbations are, similarly, an averaging of epochs of EEG data that are time-locked to an event but have been Fourier transformed to reveal the power spectral density of the signal (Makeig, 1993). EEG walking data are often displayed in event-related spectral perturbation graphs which typically cover a gait cycle, from one heel strike to the next, and present the changes in spectral power at frequencies of interest. Spectral power is often illustrated by a gradient of colors. In the processing of EEG data, it is important to select an appopriate baseline to isolate the oscillations of interest. Choosing a baseline state for comparison influences conclusions about whether there is an increase or decrease in synchronization or desynchronization. Different scientific questions require different baseline comparisons. The high temporal resolution of EEG makes it an ideal modality for studying connectivity during human locomotion. With inclusion of electromyography measurements, it is possible to combine EEG metrics and EMG metrics to assess corticomuscolar coherence during walking (Artoni et al., 2017; Petersen et al., 2012; Winslow et al., 2016). Corticomuscolar coherence can help provide insight into flow of motor commands and sensory feedback from the periphery. It has been examined the link between brain and muscle including direction of information flow during walking in multiple participants. Brain-to-muscle connectivity was stronger than muscle-to-brain connectivity and motor regions had a stronger causal influence on leg muscle activity than the non-motor regions, demonstrating the supraspinal involvement in human locomotion (Artoni et al., 2017). They also found that connectivity was strongest for distal muscles of the swing leg, which suggests that cortical control is important for ankle dorsiflexion and correct foot placement.

2.2 Machine learning for EEG

Machine learning is a branch of artificial intelligence that focuses on developing algorithms capable of learning from data and making predictions or decisions autonomously. The primary goal of this field is to build models that can generalize effectively, meaning they can make accurate predictions on previously unseen data. To achieve this, the available dataset is divided into two subsets: one used for training the model, the training set, and one used to evaluate its performance, the test set. The model's ability to generalize, i.e., to make correct predictions on new, unseen data, is a crucial measure of its effectiveness. The generalization error represents the discrepancy between the model's predictions and the true values in the test set, reflecting how well the model applies what it has learned to new data. The effectiveness of an ML algorithm is determined by its ability to minimize training error, and reduce the gap between training error and test (or generalization) error. When a classifier achieves both of these objectives, it is considered to be operating at its optimal point, as it is shown in Figure 6.

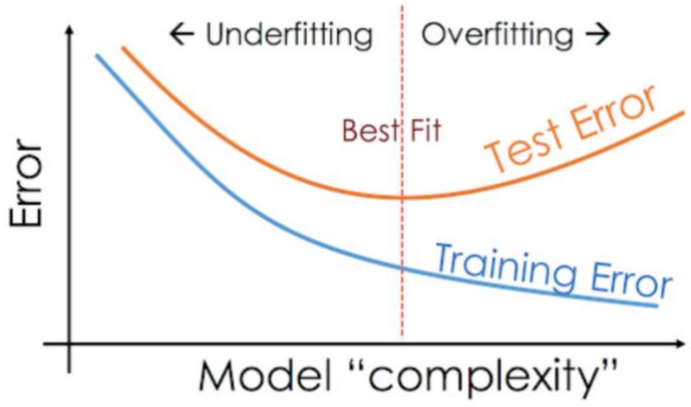

Figura 6: ML model complexity vs Error

Moving leftward from this point indicates underfitting, which occurs when the model cannot achieve a sufficiently low error on the training set. Shifting to the right, however, leads to overfitting, a condition in which the gap between training error and test error becomes too large, signaling that the model is too closely fitted to the training data and may not generalize well to new data.

In Figure7 it is represented the graphical representation of a generic EEG dataset. For each subject, the dataset can be thought of as a tensor **E** (a multi-dimensional matrix) with dimensions n times c times t, where n is the number of trials, c is the number of channels, and t is the number of time samples. The data from an individual trial, however, is represented by **E** as a simple matrix of dimensions c×t.

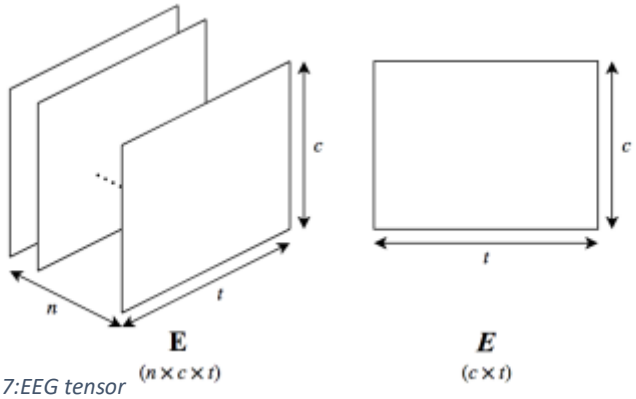

Figura 7:EEG tensor

In this thesis, several machine learning models were used, each with specific characteristics that make it suitable for particular types of problems and data. The goal of this selection was to test the performance and versatility of each algorithm, exploring how different approaches can influence the results. The models employed are: Support Vector Regression (SVR), Random Forest (RF), LightGBM, XGBoost, K-Nearest Neighbors (KNN), and Gradient Boosting Regression. Each model was chosen for its unique features, which are briefly described below. Starting with Support Vector Regression (SVR)**,** a variant of Support Vector Machine (SVM), this model stands out for the use of kernels, which allow it to handle nonlinear relationships in the data. SVR aims to find a function that minimizes the error while keeping this error within a defined margin of tolerance, focusing on the support vectors, which are the most significant points for defining the function. This approach is particularly useful in the presence of complex data but can be computationally expensive with large datasets. Next, we examine the models based on ensemble learning techniques, such as Random Forest (RF) and Gradient Boosting Regression. Random Forest constructs numerous decision trees on random subsets of data and features, and the final prediction is the average of all the trees' outputs. This approach reduces the risk of overfitting, making the model robust and versatile, especially in the presence of noisy data. However, it can become computationally costly with large datasets. On the other hand, Gradient Boosting Regression builds trees sequentially, with each new tree attempting to correct the errors of the previous model, progressively reducing the loss function. This boosting process makes the model highly accurate and suitable for complex tasks, but it requires careful parameter management to avoid overfitting. Similarly, LightGBM and XGBoost are optimized variants of Gradient Boosting, designed to improve computational efficiency, particularly with large datasets. LightGBM uses a leaf-wise growth technique for trees, allowing faster construction and deeper trees with lower computational costs, making it highly scalable and efficient even with large datasets. XGBoost, on the other hand, stands out for its use of advanced optimization techniques,

such as L1 and L2 regularization, tree pruning, and parallelization, which improve both speed and model accuracy, making it one of the most widely used algorithms in data science competitions. Finally, the K-Nearest Neighbors (KNN) model represents a completely different approach. KNN is a non-parametric model, meaning it does not require explicit training. Instead, for each new prediction, KNN finds the k nearest neighbors to the input point in the dataset and predicts a value based on the average output values of these neighbors. While simple to implement and versatile, KNN suffers from computational limitations: for each prediction, the model must calculate the distance between the input point and every other data point in the dataset, which makes it inefficient with large datasets. Additionally, KNN may suffer from the curse of dimensionality, where an increase in the number of features makes the distances between points less meaningful, reducing the model's effectiveness. In the next chapter, the materials and methods used in this research will be described, outlining how the thesis work was carried out, the data collection and preprocessing steps, and the techniques used for applying and evaluating the machine learning models mentioned above.

3 Materials and Methods

3.1 EEG and kinematics dataset

The study utilized an existing dataset of healthy participants, free from neurological disorders, who walked on a treadmill. EEG data were recorded with a 60-channel system, while six joint angles (covering bilateral hip, knee, and ankle joints) were captured via goniometric measurements, along with four EOG channels. The EOG channels served as inputs to a real-time H-infinity filter to remove eye movement artifacts. Previous studies have shown that, with meticulous setup, movement artifacts in EEG are negligible at slower speeds. Given the treadmill speed of 0.45 m/s in this study, no additional EEG artifact correction specific to walking was required. This speed was also selected with future gait rehabilitation applications in mind, as initial stages of therapy are likely to involve slow walking. Each participant completed three sessions, each comprising a 20-minute walking period followed by a 4-minute rest.

Firstly, I tried to predict kinematics, just using the EEG signals. The codes were implemented in Python, beginning with the import of EEG and joint kinematic data from two separate .txt files. From the EEG file, only the rows and columns corresponding to the time points and the 60 EEG channels were imported, while from the joint angle file, only the rows and columns representing time points and the joint angles measured through goniometry were selected. Additionally, the EEG data were normalized using the "StandardScaler" class from scikit-learn library, adjusting them to have a mean of 0 and a standard deviation of 1. This normalization step was performed to enhance the efficiency and accuracy of machine learning algorithms that would later be used for predicting joint kinematics and classifying gait phases into swing and stance. Subsequently, both the EEG and joint angle datasets were split into training and testing sets, with 90% allocated to training and 10% to testing, in preparation for future machine learning model applications. To implement them, it was essential to segment the data into temporal windows. A window length of 500 ms was selected since it is an optimale tradeoff between temporal resolution and signal robustness: it aligns with the temporal scale of locomotor dynamics (approximately half a gait cycle) and provides meaningful information while minimizing noise. An overlap of 100 ms between consecutive windows was chosen to ensure balance between temporal continuity among windows, temporal resolution and informational redundancy. Given the sampling rate of 100 Hz, the number of samples per window and the number of overlapping samples between adjacent windows were computed accordingly. This approach ensured appropriate data segmentation to optimize the model’s ability to capture relevant temporal patterns.

As it is described in 2.2, a 3D tensor was then constructed for the EEG data, with dimensions corresponding to number of EEG windows times number of samples per window times number of channels, for both the training and test sets. Similarly, a 2D matrix was created for the joint angle data, organized as number of windows times number of joint angles, for both the training and test sets. This structured arrangement of data enabled efficient handling and processing for subsequent machine learning applications. Afterwords, a loop for was executed across the number of windows in both the training and test sets to calculate the start and end indices of each window. For each window, the EEG data were stored from the first to the last sample across all channels, while only the final sample of each of the six joint angles was saved for each window. This approach was taken to align with the objective of predicting the last kinematic sample based on the corresponding EEG window. To prepare the data for machine learning model application, we extracted relevant EEG features by reshaping the 3D tensor into a 2D matrix with dimensions number of windows times number of features, where the features were derived as number of samples times number of channels. This transformation was applied to both the training and test sets. Moreover, two empty matrices with dimensions number of windows times number of joint angles were initialized to store the joint angle predictions generated by the machine learning models. The above setup was designed to maintain a one-to-one correspondence between windows and joint angles. Specifically, for each of the six joint angles, there was a direct mapping between the number of windows and the number of predictions, as only the final kinematic sample was predicted for each window. To reduce computational complexity, we selected a subset of two thousand windows for the training set and three hundred windows for the test set. This preprocessing optimized the dataset for efficient model training and evaluation.

3.2 Machine learning for movement classification

Following, machine learning models were applied to data. Each model was trained on the training set and used to predict joint angles on the test set across all six target angles. Specifically, Support Vector Regressor (SVR), Random Forest (RF), LightGBM, and XGBoost models were implemented. For the Random Forest model, a hyperparameter optimization was conducted to enhance predictive accuracy. This comprehensive modeling approach allowed for performance comparisons across algorithms in predicting joint kinematics. As mentioned previously, the initial kinematic prediction was made using only EEG decoding. However, we anticipated that this approach would not yield satisfactory results, a hypothesis that was confirmed. Consequently, we

decided to incorporate N joint angle samples as additional input into the machine learning models to enhance the accuracy of the predictions. In this approach, after loading the data, splitting into training and testing sets, and creating time windows, a 3D tensor was generated for the EEG data with the same dimensions as previously. For the joint angle kinematics, however, a 3D tensor was constructed with dimensions number of windows times number of samples times number of joint angles. Since this method leveraged both EEG data and the first N kinematic samples within each window for prediction, the EEG features were concatenated with these initial joint angle values. This enriched feature set provided a more informative basis for the machine learning models to accurately predict the final joint kinematics within each window. In this approach, SVR, Random Forest, Gradient Boosting Regressor, and KNeighbors Regressor were applied. A subset of one thousand windows from the training set was used for model fitting and prediction, optimizing computational efficiency while maintaining robust model evaluation**.** After initially addressing the kinematics decoding using only EEG signals and subsequently integrating kinematic data, the focus shifted to classifying the stance and swing phases of gait, using only EEG data. Accurately distinguishing between these phases is crucial for optimizing rehabilitative and assistive interventions. In particular, classifying stance and swing is essential for clinical applications, such as for patients using exoskeletons or those requiring gait assistance. Analyzing these phases enables more precise and targeted gait monitoring, which can be used to personalize rehabilitation treatments and enhance the effectiveness of assistive technologies. In the code implementation, for simplicity, only the peaks in the kinematic signal of the right hip angle were used to define each gait cycle. Each step cycle was defined as the segment of the kinematic signal between two consecutive peaks, with only peaks separated by at least 140 samples considered valid for a complete step. This choice was made based on a sampling frequency of 100 Hz and the fact that the typical duration of a step is approximately 1.4 seconds, meaning that a valid step cycle must consist of at least 140 samples. As it is shown in Figure 8, to distinguish between swing and stance phases, the minimum joint angle value within each step was identified.

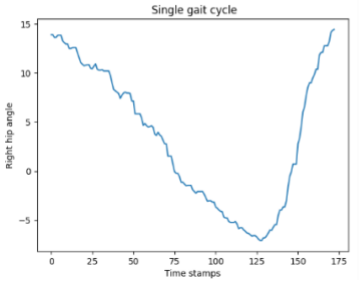

Figura 8: A complete gait cycle

The period from the start of each step to the time index of this minimum value was defined as the swing phase, while the remainder of the step cycle was designated as the stance phase. The swing phase begins at the maximum peak of the hip joint angle and ends at the subsequent minimum, during which the leg is lifted off the ground and prepared to move forward. The stance phase starts from this minimum and continues until the next maximum peak, when the foot is in contact with the ground and the hip joint supports the body's weight while it moves forward. For labeling, the swing phase vector was filled with zeros, and the stance phase vector was filled with ones. Consecutive sequences of zeros and ones were used to generate windows corresponding to the swing and stance phases, and the EEG data associated with each window were stored. Additionally, to ensure uniform EEG window lengths between swing and stance phases, the minimum window length was identified and applied as a standard length; all other windows were trimmed accordingly.

The EEG windows corresponding to the swing and stance phases were concatenated, and a target vector was created by combining a sequence of zeros with a sequence of ones of the same length. This process transformed the initial series of zeros and ones—representing individual samples during the swing and stance phases—into a format where each series of zeros corresponded to a single zero, and each series of ones corresponded to a single one. For each original series of zeros and ones, the associated EEG data was stored, effectively mapping the swing and stance phases to distinct labels for further analysis. In this case, the concatenated EEG tensor created earlier was reshaped into a two-dimensional matrix with dimensions corresponding to the number of windows times the number of features. The number of features was calculated as the product of the number of samples and the number of channels. Next, the dataset was split into training and testing sets, allocating 80% of the data for training and 20% for testing. To enhance the performance and accuracy of the machine learning models, the EEG data was normalized to have a mean of zero and a standard deviation of one. This preprocessing step ensured that the features were on a comparable scale, which is crucial for achieving optimal model performance. At this stage, machine learning models were applied to predict gait phases, swing and stance. The models employed included KNeighborsClassifier, Random Forest, and Support Vector Classifier (SVC). To evaluate the performance and accuracy of these classifiers, a confusion matrix was computed for each model, providing insights into their classification capabilities. Furthermore, the Receiver Operating Characteristic (ROC) curve was generated to assess and compare the models' ability to distinguish between the two gait phases.

The ROC curve (Receiver Operating Characteristic curve) is a graphical representation used to evaluate the performance of binary classification models. It illustrates the trade-off between the True Positive Rate (TPR), also known as sensitivity, on the y-axis, and the False Positive Rate (FPR), on the x-axis, across different classification thresholds.

An ideal classifier has a curve that reaches the top-left corner of the plot, indicating high sensitivity with low FPR. The Area Under the Curve (AUC**)** quantifies the overall performance:

- AUC = 1.0: Perfect classification.
- AUC = 0.5: Random guessing.
- AUC < 0.5: Worse than random guessing.

The ROC curve is particularly valuable for assessing classifier performance in imbalanced datasets or when the costs of false positives and false negatives differ, allowing for the selection of an optimal decision threshold based on the problem's requirements.

Since achieving extreme precision in predicting angles values is not crucial for our purposes, we transformed the regression problem into a classification task by discretizing the joint angle. Precisely, it suffices to ensure that the prediction error remains within an acceptable margin of error (denoted as Delta Δ, though its exact value has not been defined at this stage). This approach allows us to focus on capturing the general direction or range of the angular values rather than pinpointing their exact value. Therefore, after loading the data, normalizing the EEG signals, and splitting both the EEG and joint kinematics datasets into training and test sets, the minimum and maximum joint angle values were identified, and angular values have been discretized between these two values into n classes, with n ranging among three and seven. A Random Forest model was trained on the training set and subsequently tested on its ability to predict the joint angle classes in the test set. The model's performance was evaluated using both accuracy and a confusion matrix, providing insight into its classification capabilities. Lastly, the prediction accuracy of the last sample in each window was analyzed as a function of the number of samples considered and the window size. A three-dimensional plot was generated to illustrate these results. The workflow included loading the data, normalizing the EEG signals, creating and populating time windows, and splitting the datasets into training and test sets. Each window was populated with both EEG and kinematic data, forming a three-dimensional tensor of size number of windows times number of samples times number of channels for the EEG data, and a tensor of size number of windows times number of samples times number of joint angles for the kinematic dataset.

The performance of the Random Forest model was evaluated in predicting the last joint angle sample within each window. This assessment was quantified using the Mean Square Error (MSE), providing

a robust metric to analyze the model's predictive accuracy under varying window sizes and sample configurations.

The application of machine learning techniques to predict gait kinematics and classify gait cycle phases, such as stance and swing, in healthy individuals during walking has provided valuable insights. These studies have demonstrated the potential of EEG signals to decode motor patterns in controlled settings, typically involving constant-speed walking without external support. However, such simplified scenarios do not fully encompass the complexities of real-world locomotion or the intricate interplay between neural, muscular, and biomechanical systems.

A natural progression toward a more advanced framework involves implementing experiments that integrate simultaneous recordings of EEG, EMG, and inertial measurement units (IMUs) during gait assisted by exoskeletons. Exoskeletons with adjustable levels of assistance introduce dynamic conditions that require coordination between central motor control, peripheral muscle activation, and mechanical responses. Such experiments provide an opportunity to study how EEG signals adapt to these more complex and realistic scenarios, such as varying levels of robotic assistance, and how these adaptations can inform the optimization of assistive interventions.

This experiment was further motivated by several unresolved questions in robotic-assisted gait (RAG). It remains unclear how the brain and muscles communicate during RAG walking, particularly in terms of supraspinal neural drive, and whether RAG promotes naturally occurring muscle activation patterns (muscle synergies) that could be utilized for rehabilitation. Moreover, the functional activity of the central nervous system, peripheral muscles, and biomechanics under RAG, and how these interactions change with varying levels of assistance, remain poorly understood. To address these questions, the experiment aimed to achieve a comprehensive understanding of how the neuromusculoskeletal system is orchestrated during different RAG conditions and to develop a robust framework capable of reliably measuring the interplay between the central and peripheral systems during assisted walking.

Here, we present the details of the experimental design, conducted on healthy individuals walking with an exoskeleton at different levels of robotic assistance. This protocol leverages multimodal data acquisition to investigate these open questions, laying the groundwork for future advancements in robotic rehabilitation technologies.

3.3 Participants

A total of seven able-bodied subjects were recruited for the study.

Participant	Gender	Age
S0001	Male	24
S0002	Female	22
S0003	Female	23
S0004	Male	23
S0005	Female	22
S0006	Female	22
S0007	Female	19

Table 1: Participants

The experimental setup consisted of an EEG acquisition system, an EMG acquisition system, IMUs and an exoskeleton. To ensure proper skin impedance, skin preparation was performed first both for EEG and EMG, followed by the application of a conductive gel using syringes just for EEG. EEG recordings were obtained using a 64 channel EEG system (ANT Neuro eegoTMmylab, the Netherlands), with different cap sizes selected based on the head circumference of the participant. The choosen sampfling frequency was 1000Hz (Seo et al., 2015; Zhang et al., 2016). Ag-AgCl electrodes were positioned on the scalp following the 10/20 international standard positioning system, with CPz serving as the reference channel and a ground electrode in correspondence of AFz. The impedance of the electrodes was maintained below 50 kΩ. For the EMG acquisition, 14 sensors were employed, alongside an amplifier, the LabJack T7-Pro which is the data acquisition device, a Mux80 that is an analog input expansion board that adds 80 analog input channels and Ethernet Cable to transfer EMG and IMU data acquired by one computer to another one used for EEG acquisition. As for the IMU setup, an Awinda Station was used in conjunction with a BNC cable, an Arduino for triggers synchronization and 8 sensors attached to the lower limb. As for the exoskeleton, only the software to control its level of assistance was needed.

3.4 Experimental setup

The experimental protocol was structured in several phases, beginning with the placement of EMG electrodes, followed by the IMU sensors, then the EEG cap, and finally the exoskeleton. To enhance efficiency and ease of execution throughout the experiment, Graphical User Interfaces (GUIs) were developed using the PyQt5 library in Python.

During the experiment, participants, after being equipped with all the electrodes and sensors for the various signals to be acquired, walked on a treadmill while performing different tasks. Unfortunately, the experiments were conducted in a room that was not magnetically or electrically isolated. Therefore, the quality of the signals was frequently monitored throughout the various stages of the experiment by observing the physiological signals and IMU data in real-time. Before starting the electrode placement phase, participants were asked to provide general information. By pressing a specific button in the GUI, an interface appeared where the operator could enter participant details such as first name, last name, date of birth, gender, and the date of the experiment. The "SAVE" button allowed the data to be stored in a .txt file.

3.4.1 EMG sensors placement

After performing skin preparation to clean the area, the electrodes for EMG acquisition were positioned. In this experiment, the selected muscles for both legs were: the Gluteus Maximus, Rectus Femoris, Vastus Lateralis, Biceps Femoris, Tibialis Anterior, Medial Gastrocnemius, and Soleus. The sensors were connected to the electrode patches, and the Mux80 was directly connected to the LabJack T7-Pro via the DB37 connector.

3.4.2 IMU sensors placement

The selected IMU sensors were: Stern, and then, for both legs, Pelvis, Upper leg, lower leg and foot. They have been all placed with the light on the down left corner in order to have a good 3D reconstruction of the body of the participant on the Xsens software.

3.4.3 EEG cap placement

The cap for EEG was placed on the participants' head after being cleaned with an alcohol solution, ensuring that Cz electrode was positioned at the vertex. Highly-conductive electrolyte gel was injected into the hole of each electrode using a specific syringe, and the impedance values were verified and adjusted by referring to the interface on the EEG system. During EEG cap placing, participants sat comfortably on a chair positioned in front of a screen, and they were instructed to relax and avoid body and facial movements to make it easy placing the cap and adapt the impedance.

3.5 Experimental protocol

Following the placement of electrodes for signal acquisition, the participant was provided with a detailed briefing regarding the procedures to be performed. The initial phase of the protocol involved a series of resting tasks. The participant was required to stand for two minutes with their eyes closed, followed by an additional two minutes with their eyes open. The order of execution of these two conditions was randomized and varied from subject to subject. Subsequently, the walking tasks commenced, initially without any technological support. The first task involved walking on a treadmill at a speed of 3 km/h (Presacco et al., 2011). From the second task onwards, the participant wore an exoskeleton. The second task was dedicated to familiarization with the exoskeleton, involving a 10-minute walk at the same speed of 3 km/h on the treadmill. The following phases included three distinct tasks, all performed with the exoskeleton. The third task involved walking for 6 minutes at 3 km/h with the exoskeleton set to a high assistance level (100%). The fourth task required walking for 6 minutes at the same speed with the exoskeleton set to a low assistance level (80%). Finally, the fifth task involved walking for 6 minutes at 3 km/h with the exoskeleton in passive mode. At the end of these tasks, the exoskeleton was removed, and the participant was required to repeat the two initial resting conditions for an additional period of two minutes each.

3.6 GUI implementation

The EEG GUI, accessible by the command prompt of the computer used to acquire EEG data, was designed to streamline the procedure. After the gel injection,to ensure that the impedance had been properly adjusted, it was necessary to click the "Check Impedance" button in the GUI (Figure 9).

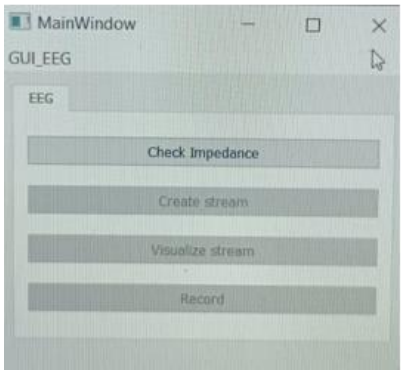

Figure 9: GUI

Subsequently, a window would appear, allowing for the selection of the headset worn by the participant and the specification of the maximum tolerated impedance threshold. This window also provided the option to view the impedance value for each individual channel (Figure 10).

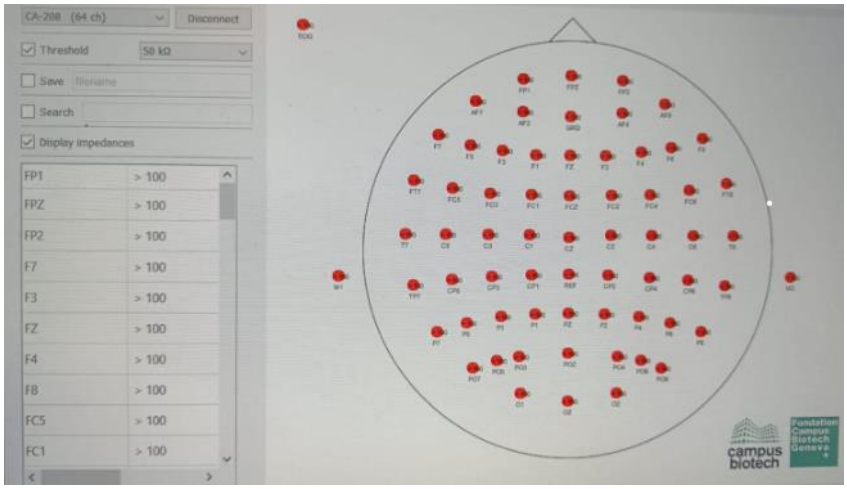

Figure 10: GUI "Check impedance"

After saving the impedance file and closing the window, the "Create stream" button in the GUI would become active (Figure 11). Upon clicking this button, it was necessary to select the headset used for the EEG recording and specify the sampling rate, 1000 Hz in our case (Figure 12).

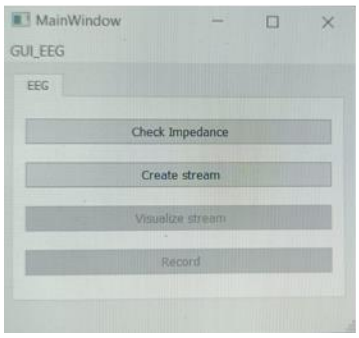

Figure 11: "Create stream" button enabled

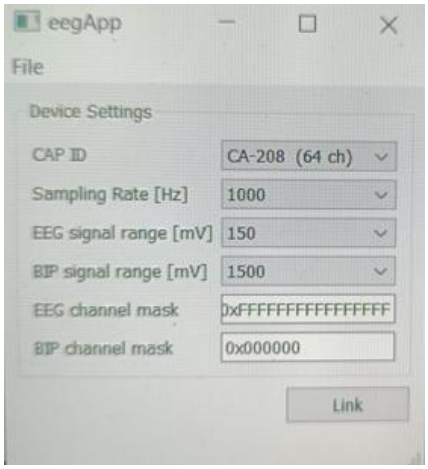

Figure 12: Stream creation

Subsequently, the "Visualize stream" and "Record" buttons were enabled. Clicking "Visualize stream" allowed for real-time viewing of the participant's EEG, providing the opportunity to check for any interference or noisy channels. By clicking "Record", the Lab Recorder window would open, where the specific streams to be recorded could be selected and a name for the saved file could be specified (Figure 13). The data recording process would then begin by clicking "Start" (Figure 14).

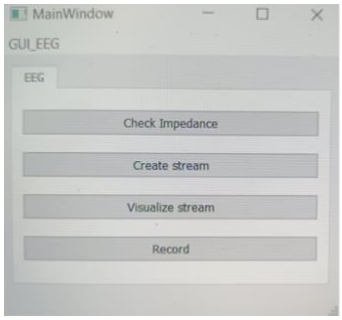

Figure 13: "Visualize stream" and "Record" buttons enabled

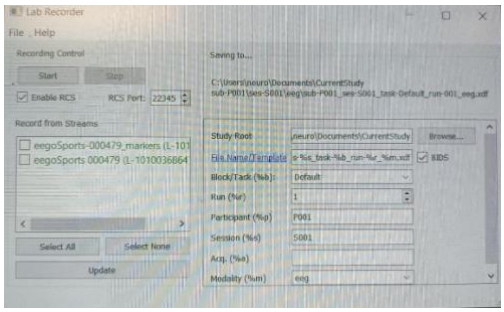

Figure 14:Start recordings

3.7 LSL protocol for synchronization

The experimental setup ensures precise synchronization of data from multiple sources, including EEG, EMG, IMUs, and exoskeleton sensors, using a well-coordinated pipeline. The process begins with the X-Sens software, which sends synchronization triggers at the start and end of each trial. These triggers are captured by the Lab Recorder via Synch Streams created by the main application. The workflow is structured such that Lab Recorder is started first, followed by the recording of exoskeleton data, and then the initiation of X-Sens to ensure all data streams are integrated into Lab Recorder. Synchronization is managed by the Awinda station, which sends a signal when "Start" is pressed in the main app. This signal is transmitted to an Arduino, which converts the voltage difference into a digital trigger. The trigger is then translated into Lab Streaming Layer (LSL) format, making it visible and recordable in Lab Recorder. This setup guarantees accurate temporal alignment of all data streams, enabling robust and synchronized analysis of neural, muscular, and biomechanical signals during the experimental trials.

4 EEG preprocessing and data analysis

After conducting experiments on patients, I carried out preprocessing of EEG data. A bandpass filter was applied to the EEG signal, specifically between 1 and 60 Hz (Artoni et al., 2017), along with a Notch filter at 50 Hz (Menicucci et al., 2014). This processing aimed to focus on the most relevant EEG components while eliminating frequencies associated with artifacts, eye movements, ambient noise, and electrical network interference.

4.1 EEG processing via EEGlab

Then I moved to EEGlab, that is an open-source software suite designed for the analysis of EEG (electroencephalographic) and MEG (magnetoencephalographic) data, developed within the MATLAB environment. Primarily used by neuroscientists and researchers, EEGLAB provides advanced tools for visualizing, processing, and statistically analyzing neurophysiological data, with

a particular emphasis on EEG recordings. The following preprocessing pipeline was implemented in EEGLAB for data acquired during the experiments. This approach was defined through iterative testing and adjustments to optimize data clarity, enhance visualization quality, and minimize noise. After several trials and errors, and according to Tortora et al., 2020, the following steps summarize the finalized protocol:

1. **Removal of Outer Circumference**: Data channels along the outer circumference were removed, resulting in a final count of 41 channels, to be focused on the sensory-motor area of the brain.
2. **Trimming of Initial and Final Data Segments**: Sections of data at the beginning and end of each recording that were deemed unreliable due to interference from the cable connecting the exoskeleton to the Awinda station, were removed.
3. **Detection of Bad Channels via Kurtosis Algorithm**: Channels with abnormal characteristics were identified using the Kurtosis algorithm,
4. **Inspection of Data Spectrum**: The data spectrum was visually checked to confirm data quality and identify any anomalies.
5. **Independent Component Analysis (ICA)**: ICA was performed to identify and remove eye movement artifacts, enhancing data accuracy.
6. **Spherical Interpolation of Bad Channels**: Channels identified as bad via Kurtosis were corrected using spherical interpolation.
7. **Re-referencing to Common Average Reference (CAR)**: Re-referencing was applied to improve signal quality.
8. **Data Saving**: The preprocessed data were saved in .mat format, including also EMG components.

This pipeline was selected as the optimal method to achieve a clean, reliable dataset suitable for subsequent analysis. It was decided to apply ICA before CAR because ICA isolates and removes artifacts (such as eye movements or muscle activity) and separates independent components, preserving neural signals. Applying CAR first could spread common noise across all channels, potentially compromising the ability of ICA to accurately separate independent sources.

Taking into consideration , for example, patient P0006 high level of exo assiatance , we proceeded as it follows (Figure 15, Figure 16):

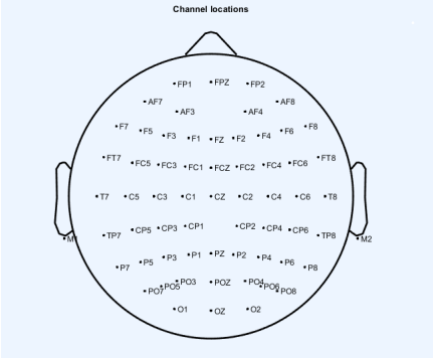

Figure 15: Head with 63 electrodes

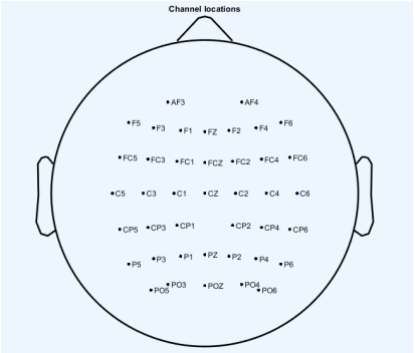

Figure 16: Head after remotion of outer circumference

As we can see, the first step has been removing the outer circumference, to be focused only on the sensory-motor area. Then, it is evident the noise due to the exo cable, and this is the reason why we had to remove bad portions of data, both at the beginning and at the end of each task of the experiment (Figure 17).

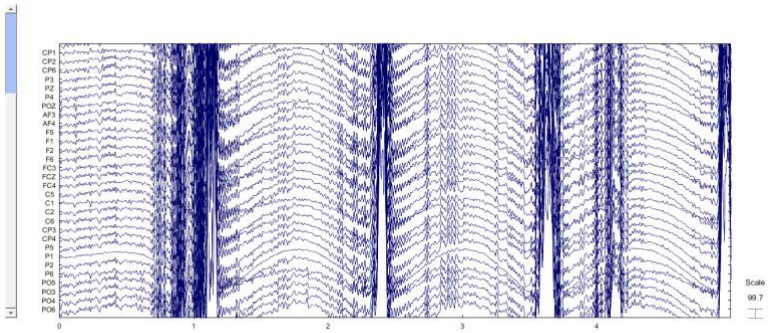

Figure 17: Noisy data due to exo cable

Then, during the experiment we detouched the exo cable, since it was necessary only for sending the trigger for the synchronization and the signal become cleaner, as it follows (Figure 18):

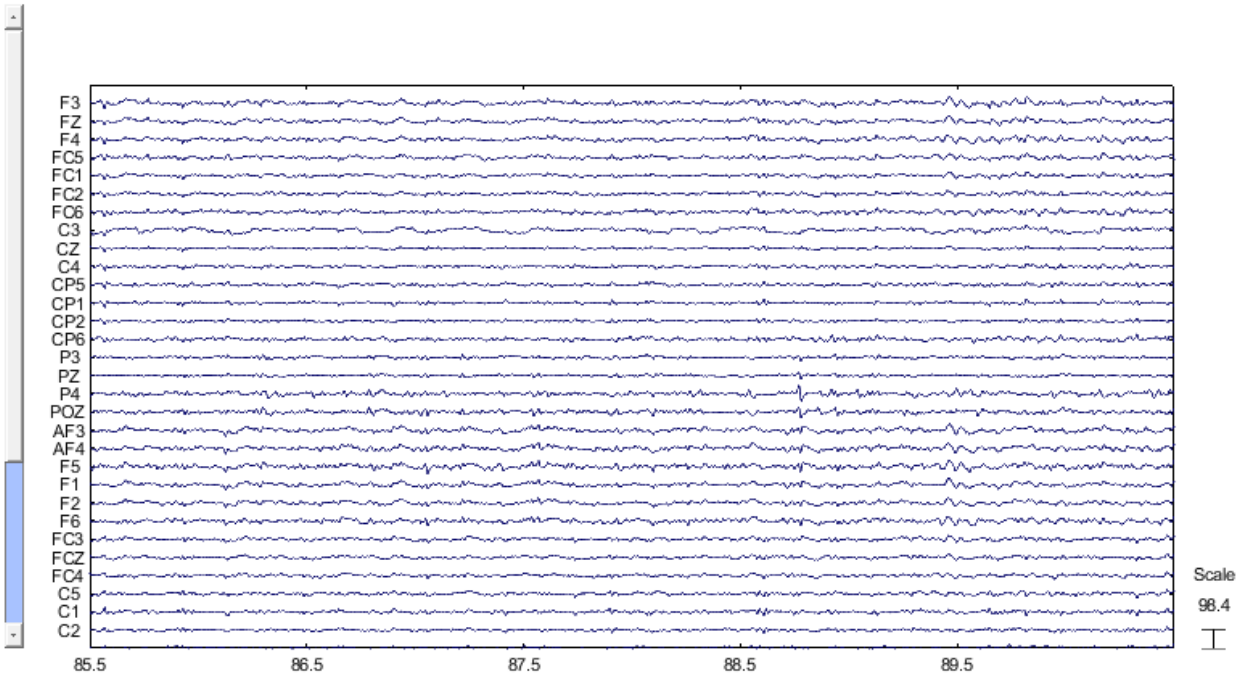

Figure 18: Cleaned Signals

Then , I removed bad portions of data, both at the beginning and at the end of the experiment (Figure 19, Figure 20).

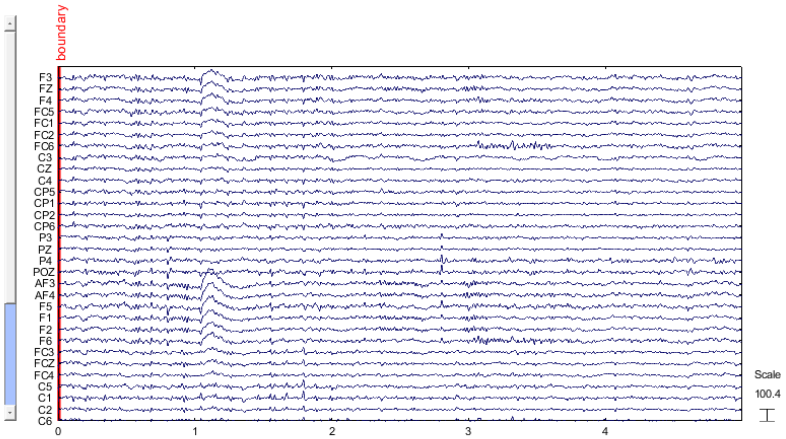

Figure 19: Remotion of bad data at the beginning

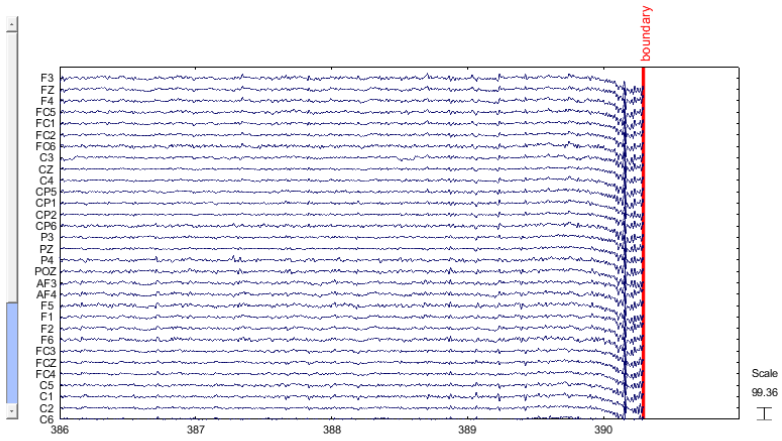

Figure 20: Remotion of bad portions of data at the end

Afterwords, together with Kurtosis algorithm, I looked at the power in the frequency domain to remove EEG channels, for two main reasons (Figure 21):

- **High-Frequency Noise**: EEG channels displaying elevated power in atypical frequency bands, particularly in higher frequencies (>30 Hz), often contain noise from external sources or non-neural artifacts, such as muscle activity (EMG) or electronic interference. These signals can interfere with analyses focusing on lower frequency ranges, where the neural components relevant to gait studies are typically found.
- **Correlation with Muscle Movements**: During gait, muscle artifacts can easily overlap with EEG signals, especially in higher frequencies. Removing noisy channels based on power in these frequency bands helps to better isolate pure brain activity from unwanted muscle signals.

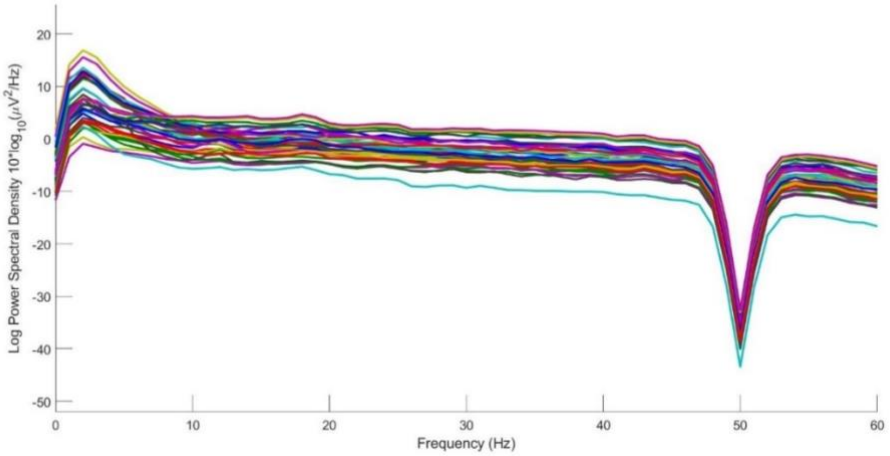

Figure 21: Power spectral density

Subsequently, I applied the ICA. ICA (Independent Component Analysis) is used in EEG and other applications to separate overlapping signals into independent components. Here's why it is applied:

1. **Artifact Removal**:
    - **Non-Neural Artifacts**: ICA is particularly effective in isolating and removing common non-neural artifacts in EEG data, such as eye movements (saccades and blinks), muscle activity (EMG), and interference from electronic devices. If not removed, these signals can overlap with brain activity, compromising the accuracy of analyses.
    - **Isolation of Specific Components**: ICA allows for the identification of independent components that represent these artifacts and facilitates their removal, keeping the brain signal clean.

These were the components before ICA (Figure 22, Figure 23): here and in the next figure it can be seen the blink components circled in blue with the typical pointed shape pointing up or down.

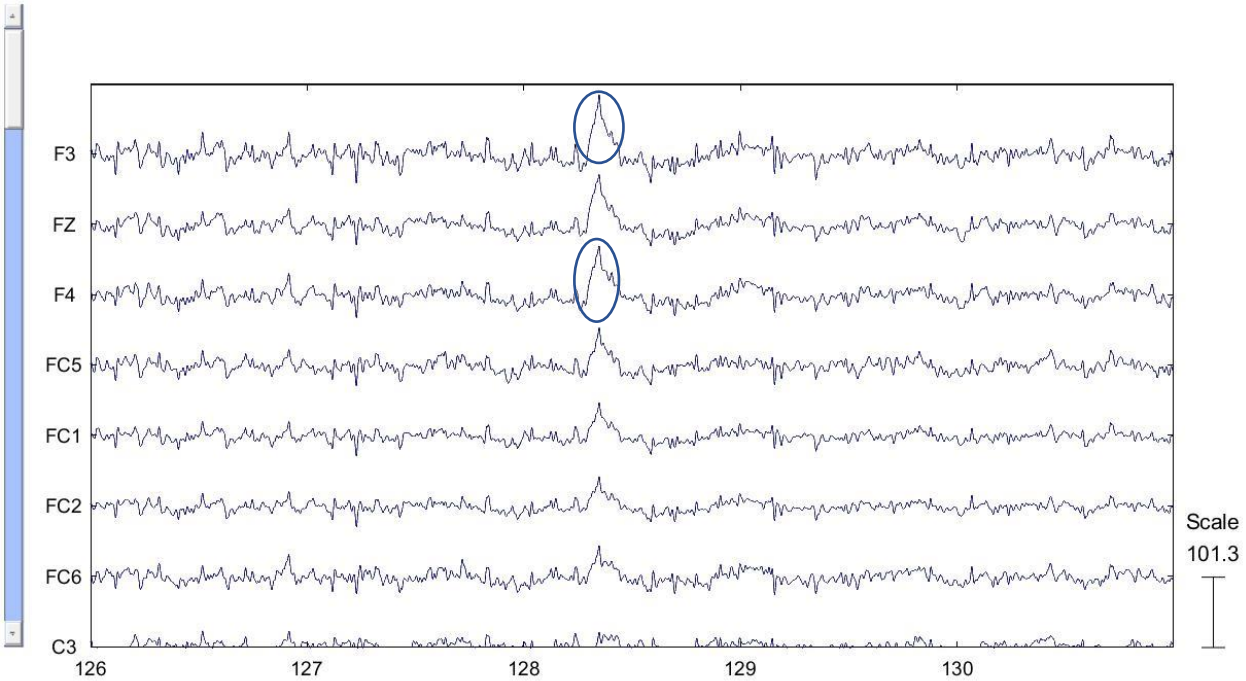

Figure 22: Eye blink

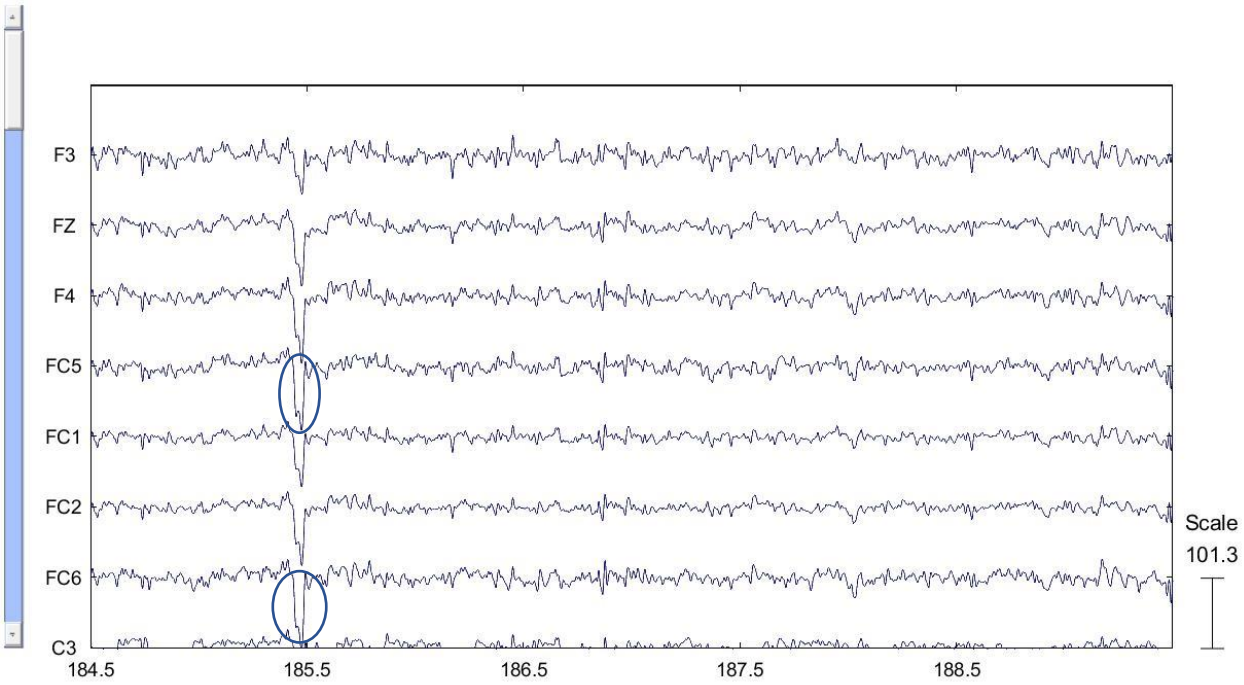

Figure 23: Eye blink

Below it is evident the eyes lateral movement circled in red (Figure 24).

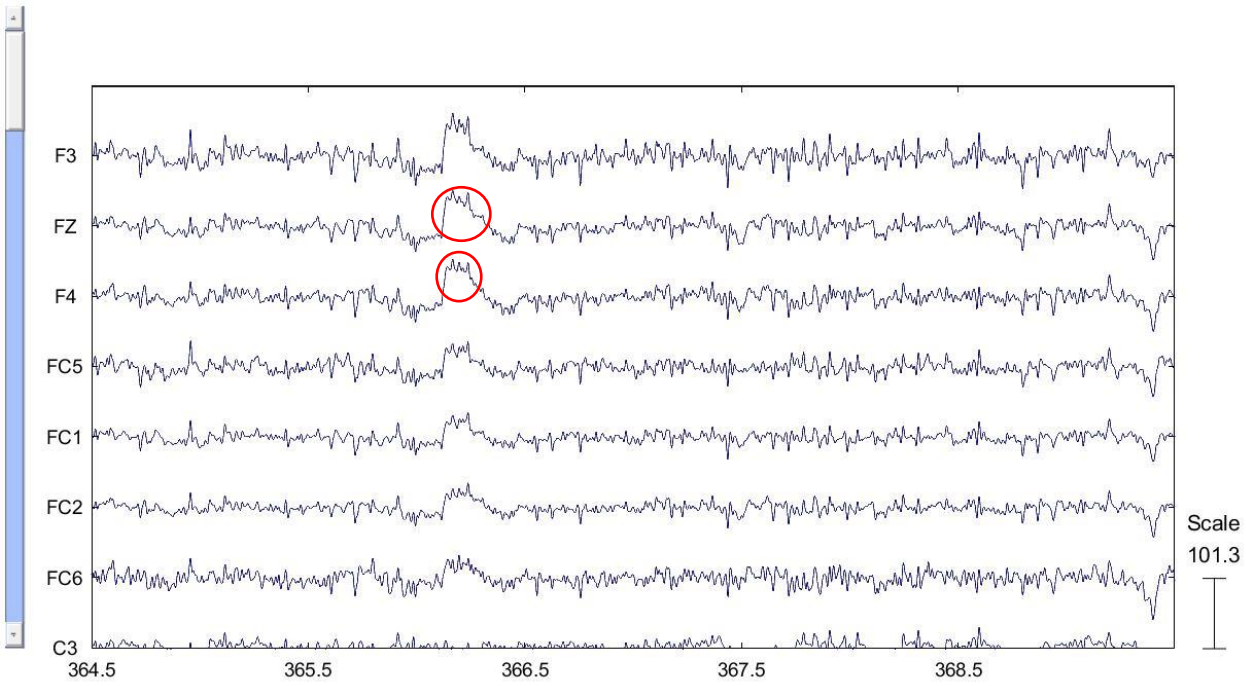

Figure 24: Eyes lateral movement

The following are the ICs, computed on EEGlab (Figure 25):

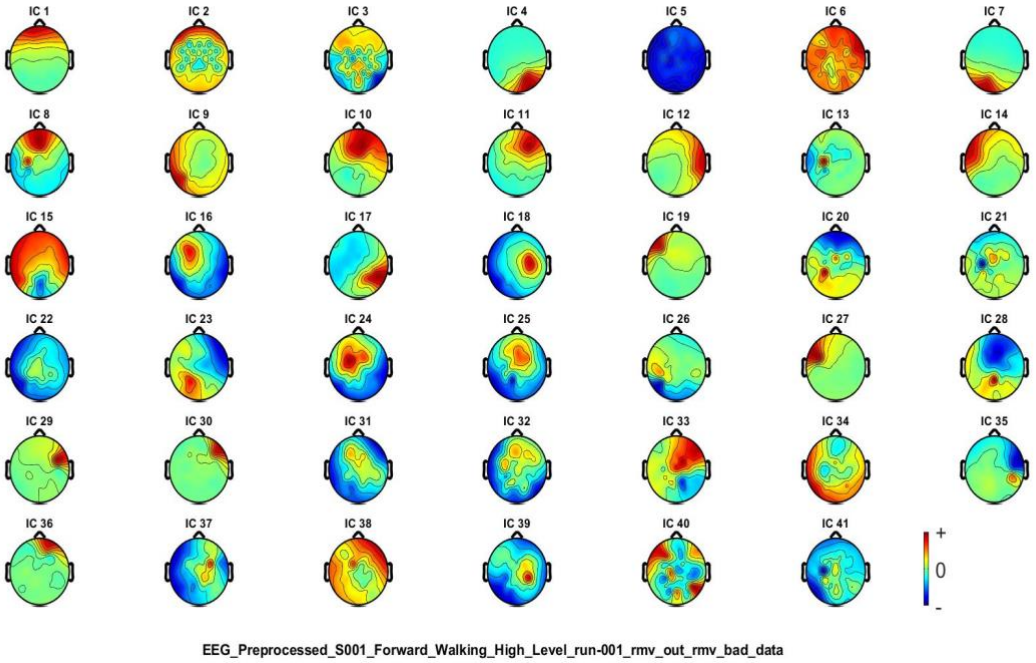

Figure 25: ICs

As it can be seen, the IC 1 is an ocular component, or blink. After detecting the indipendent components, we proceeded removing them. Indeed, now, either the signals no longer have the artifacts mentioned before or these have been greatly reduced (Figure 26, Figure 27, Figure 28).

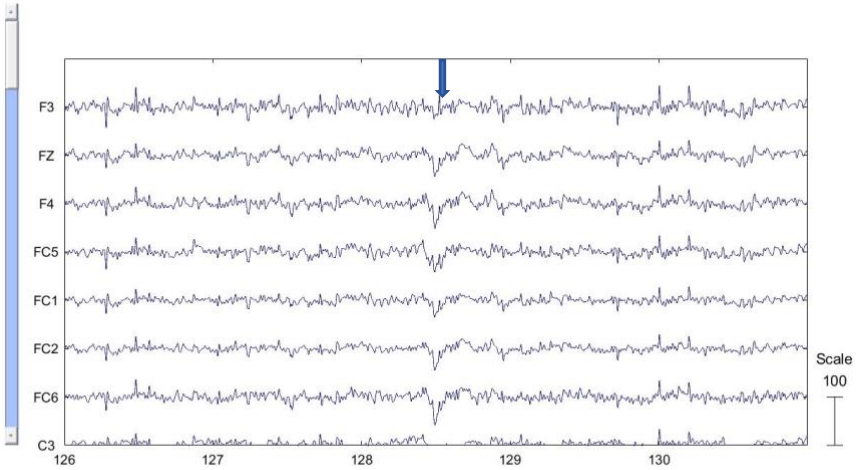

Figure 26: Remotion of eye blink component

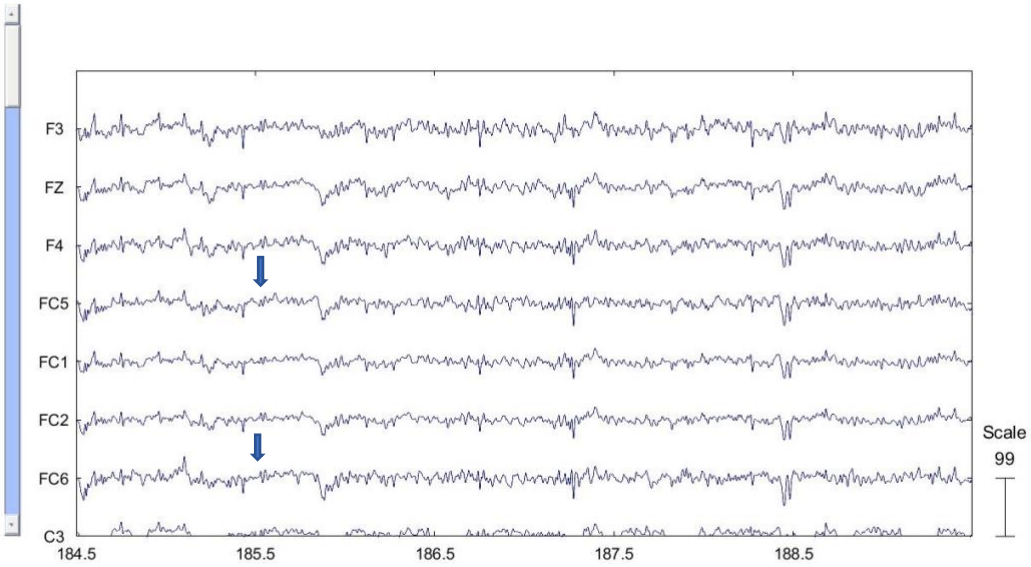

Figure 27: Remotion of eye blink component

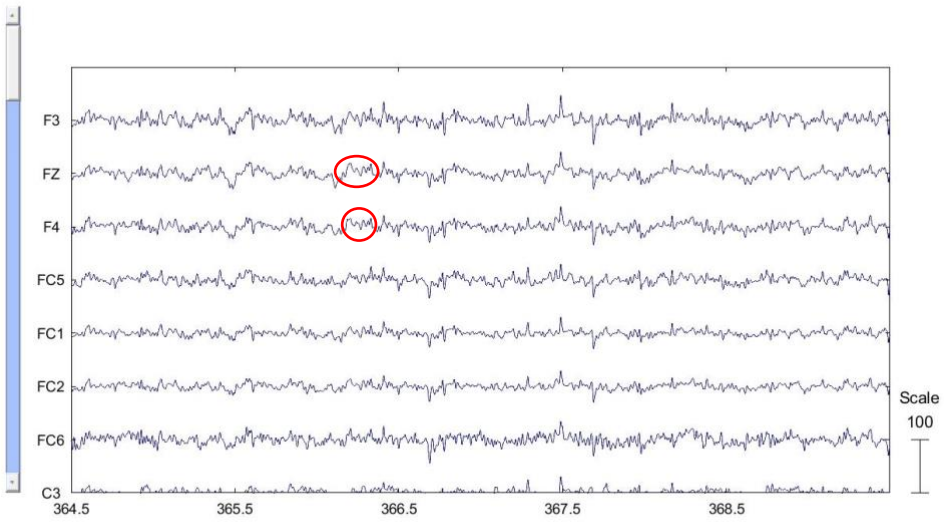

Figure 28: Remotion of eye lateral movement component

Then, we procedeed with spherical interpolation of bad channels, detected by Kurtosis algorithm and by visual inspection of power in frequency domain. In following figures it can be seen that noise on P06 is evidently reduced (Figure 29, Figure 30):

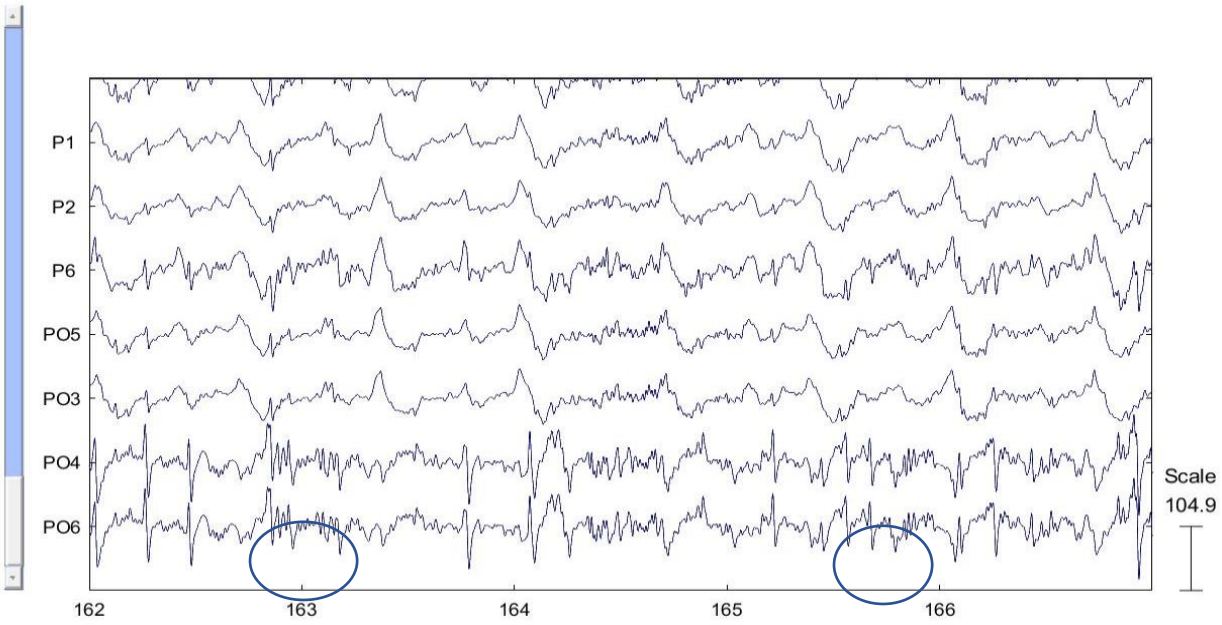

Figure 29: Noisy P06

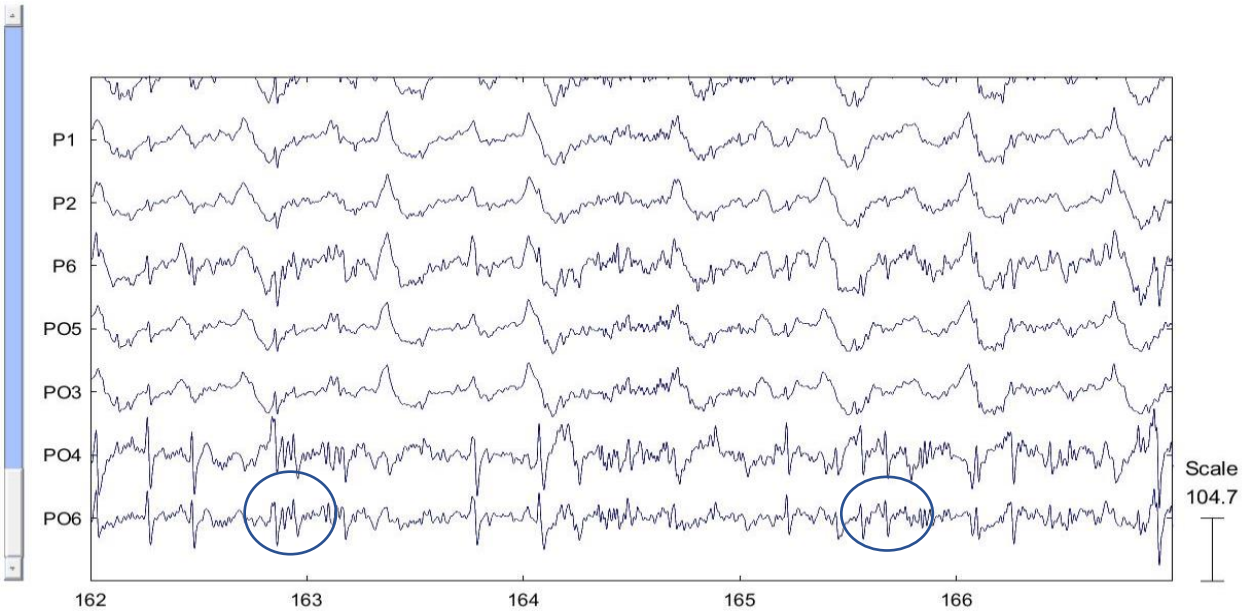

Figure 30: P06 spherically interpolated

Lastly, I applied Common Average Reference (CAR). The Common Average Reference (CAR) is a referencing technique used in the analysis of EEG (electroencephalogram) data. It is a form of

referencing where the reference signal is the average of the signals recorded from all electrodes. In other words, the signal value from each electrode is subtracted from the average of all electrode signals. This approach aims to improve the quality of EEG data and reduce noise and artifacts.

This is how the signal appears after CAR (Figure 31):

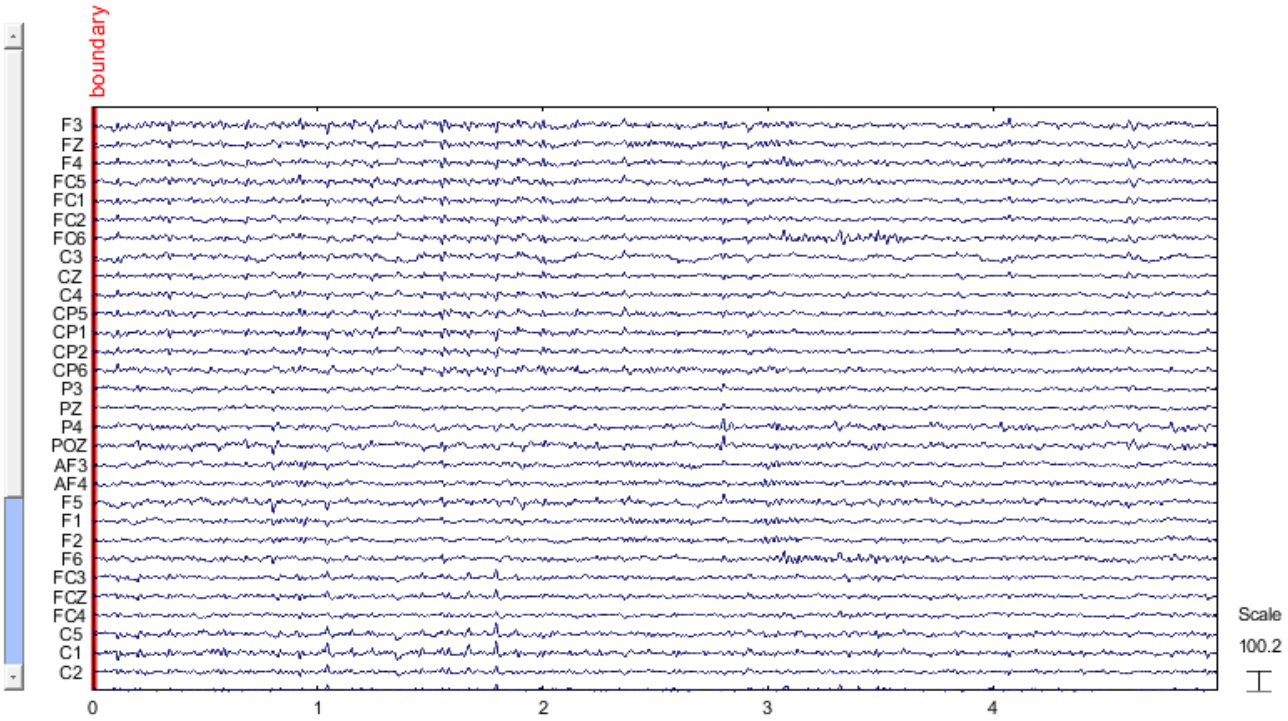

Figure 31: EEG signals after CAR

5 Results

The findings are organized to address the research objectives and are supported by graphical representations and statistical analyses where appropriate.

5.1 EEG-based predictions

In this section, we first present the kinematic prediction results based exclusively on EEG data. The performance of all four predictive models employed in the study is illustrated, providing a comprehensive evaluation and comparison of their effectiveness. Below, it is shown result of SVR model application (Figure 32, Figure 33, Figure 34, Figure 35, Figure 36, Figure 37).

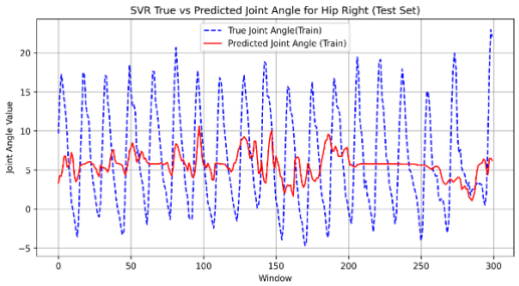

Figure 32: SVR Hip Right (Test Set)

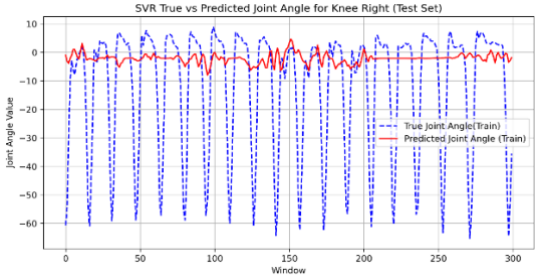

Figure 33: SVR Knee Right (Test Set)

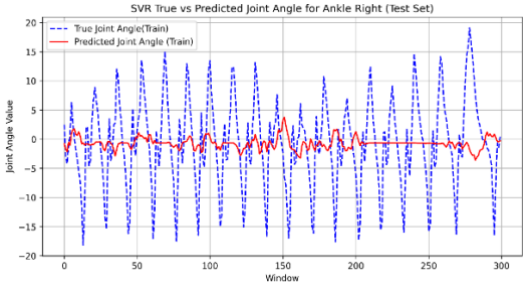

Figure 34: SVR Ankle Right (Test Set)

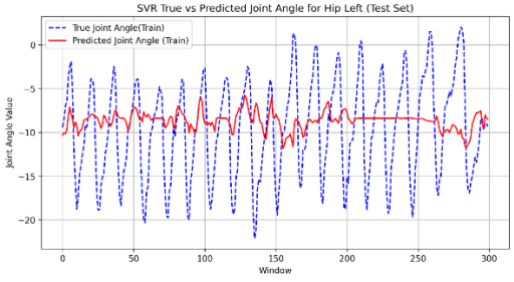

Figure 35: SVR Hip Left (Test Set)

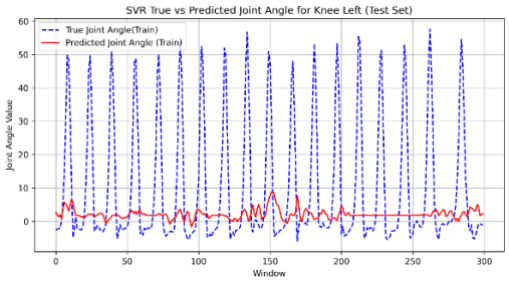

Figure 36: SVR Knee Left (Test Set)

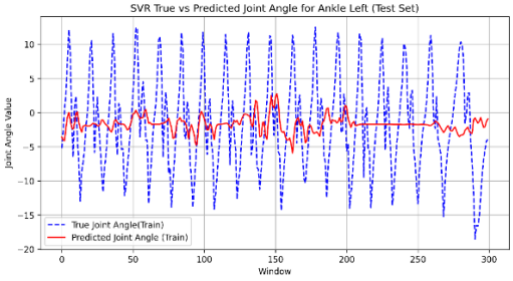

Figure 37: SVR Ankle Left (Test Set)

Below I present the results obtained from applying the Random Forest model (Figure 38, Figure 39, Figure 40, Figure 41, Figure 42, Figure 43).

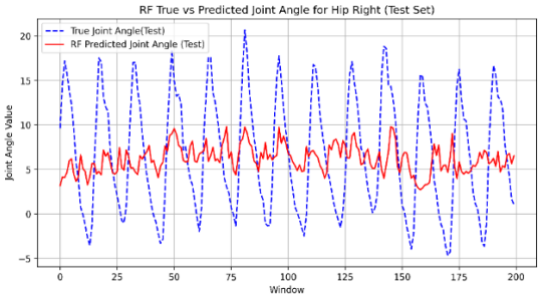

Figure 38: RF Hip Right (Test Set)

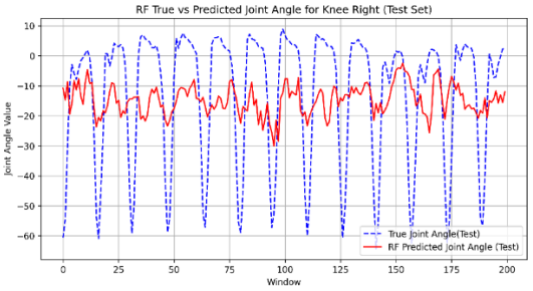

Figure 39: RF Knee Right (Test Set)

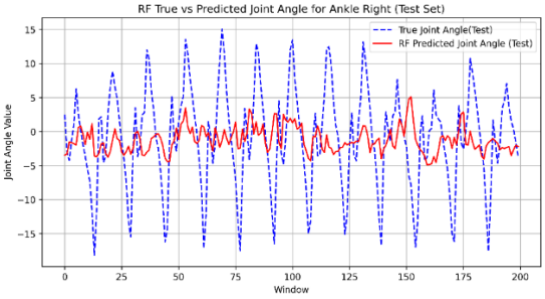

Figure 40: RF Ankle Right (Test Set)

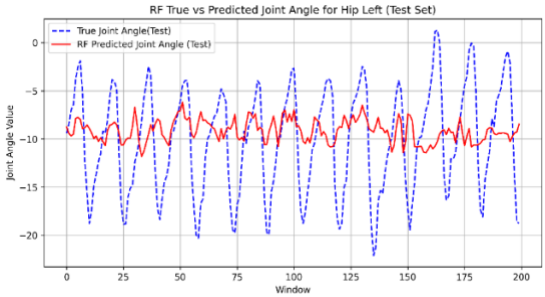

Figure 41: RF Hip Left (Test Set)

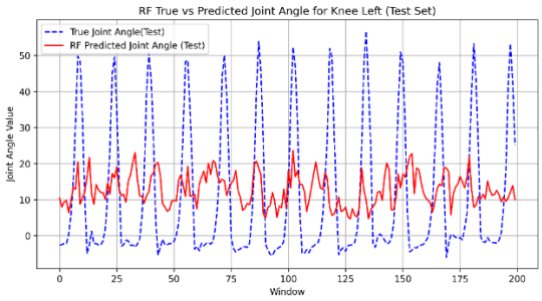

Figure 42: RF Knee Left (Test Set)

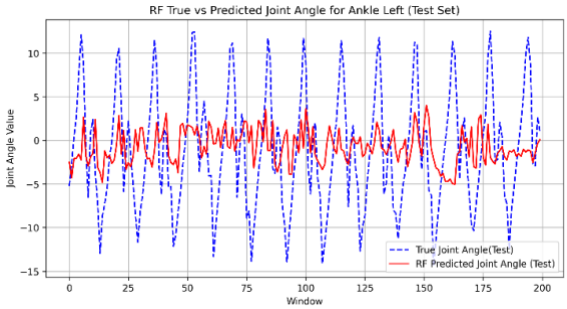

Figure 43: RF Ankle Left (Test Set)

Here the results obtained from applying the LightGbm model are shown (Figure 44, Figure 45, Figure 46, Figure 47, Figure 48, Figure 49).

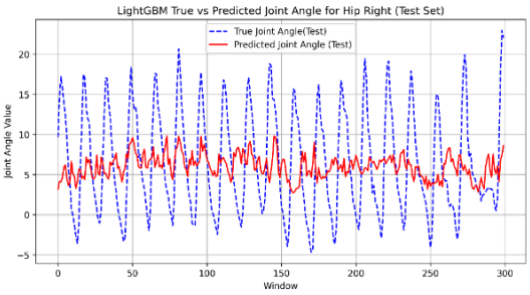

Figure 44: LightGBM Hip Right (Test Set)

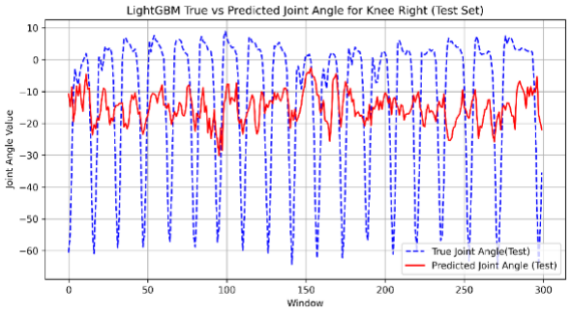

Figure 45: LightGBM Knee Right (Test Set)

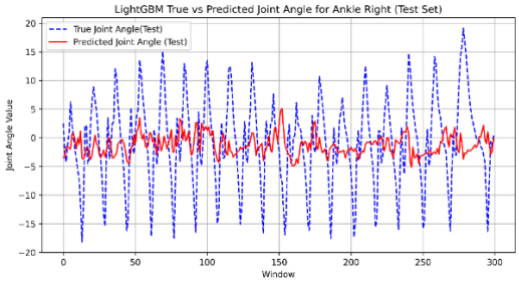

Figure 46: LightGBM Ankle Right (Test Set)

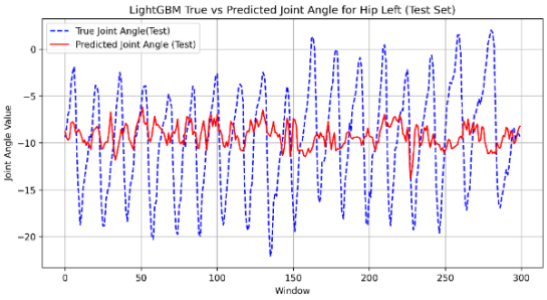

Figure 47: LightGBM Hip Left (Test Set)

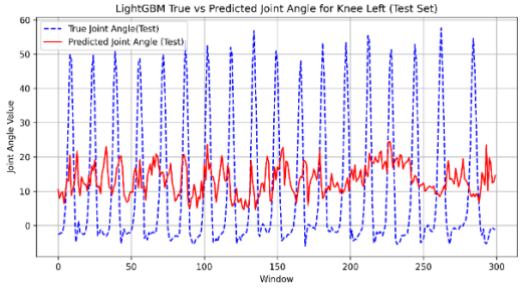

Figure 48: LightGBM Knee Left (Test Set

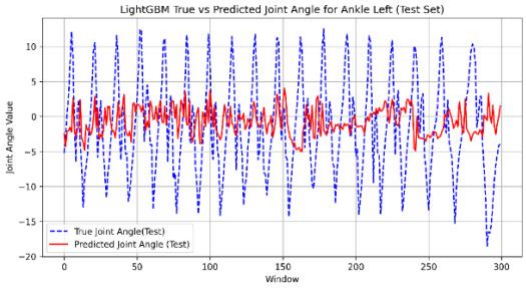

Figure 49: LightGBM Ankle Left (Test Set)

Following the results obtained from applying the XGBoost model are presented (Figure 50, Figure 51, Figure 52, Figure 53, Figure 54, Figure 55).

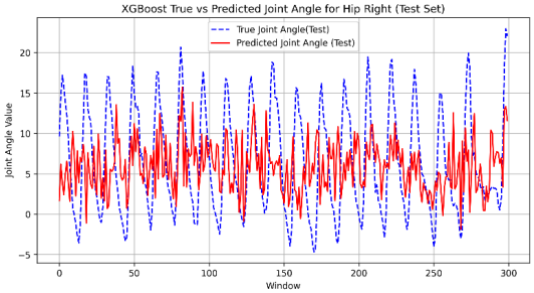

Figure 50: XGBoost Hip Right (Test Set)

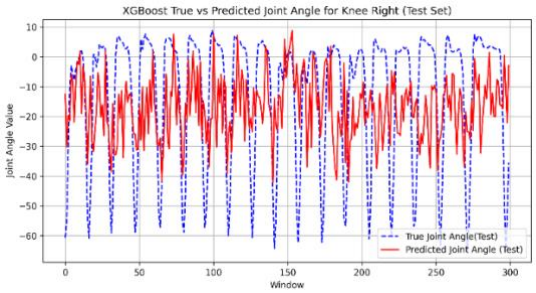

Figure 51: XGBoost Knee Right (Test Set)

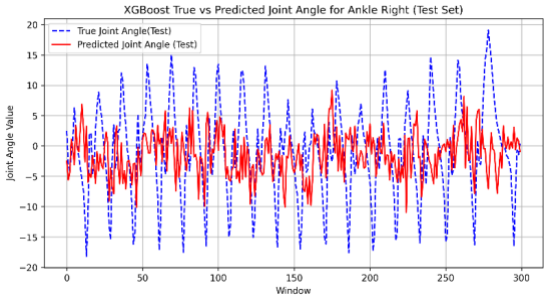

Figure 52: XGBoost Ankle Right (Test Set)

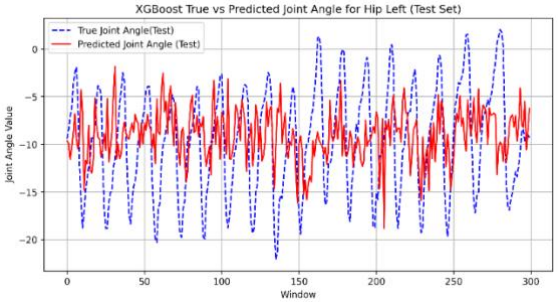

Figure 53: XGBoost Hip Left (Test Set)

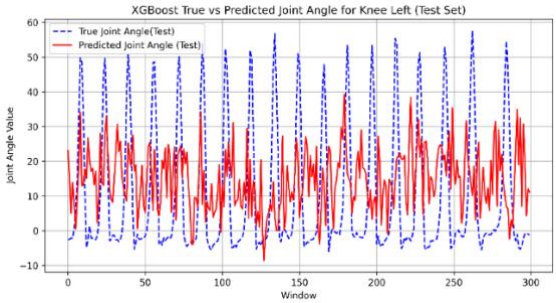

Figure 54: XGBoost Knee Left (Test Set)

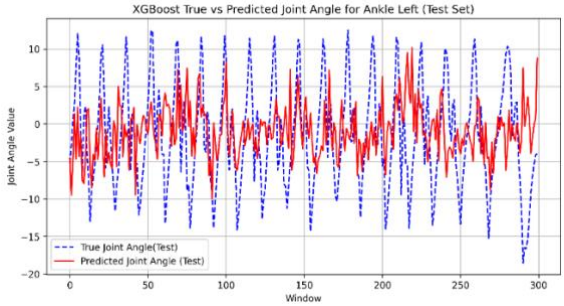

Figure 55: XGBoost Ankle Left (Test Set)

As shown by the previous graphs, EEG signals prove insufficient for accurately reconstructing gait kinematics. Based on the metrics considered (MAE, MSE, R^2) (Figure 55), it can be observed that predictions for the ankle and hip joints have the lowest error values, except for the MAE of the SVR model, likely because these two joints have the smallest range of motion during gait. In contrast, for all machine learning models and for all MSE values, as well as for all MAEs except SVR model, the errors in predicting knee joint angles are the highest. This is because the knee exhibits greater variability in joint angles, with significant changes between flexion and extension.

Model	Joint	MSE	MAE	R^2
Random Forest	Joint 1	81.75870	8.37037	-1.67992
Random Forest	Joint 2	231.25828	13.08754	-0.65218
Random Forest	Joint 3	22.40042	3.38329	-0.19631
Random Forest	Joint 4	99.83610	9.22856	-1.87569
Random Forest	Joint 5	235.74013	13.69869	-1.02074
Random Forest	Joint 6	22.90796	3.45467	-0.36400
LightGBM	Joint 1	79.55546	8.07870	-1.60770
LightGBM	Joint 2	232.89306	12.17339	-0.66386
LightGBM	Joint 3	28.08934	3.93223	-0.50013
LightGBM	Joint 4	107.15387	9.43283	-2.08647
LightGBM	Joint 5	274.16920	13.87680	-1.35015
LightGBM	Joint 6	26.65088	3.80216	-0.58686
XGBoost	Joint 1	87.18892	8.35618	-1.85791
XGBoost	Joint 2	278.81178	13.14271	-0.99191
XGBoost	Joint 3	32.31992	4.31956	-0.72606
XGBoost	Joint 4	106.60683	9.38322	-2.07071
XGBoost	Joint 5	302.30576	14.25928	-1.59133
XGBoost	Joint 6	29.27783	4.03274	-0.74328
SVR	Joint 1	78.05146	8.07516	-1.55840
SVR	Joint 2	147.41968	6.19082	-0.05321
SVR	Joint 3	21.38832	3.18051	-0.14226
SVR	Joint 4	94.66827	8.96611	-1.72684
SVR	Joint 5	121.60430	6.00206	-0.04238
SVR	Joint 6	22.20519	3.28883	-0.32215

Figure 56: Models performances

5.2 EEG and kinematics based predictions

Below the angular prediction graphs are presented, considering the first N kinematic samples (with N=4) of each window. Firstly the SVR plots are shown (Figure 57, Figure 58, Figure 59, Figure 60, Figure 61, Figure 62).

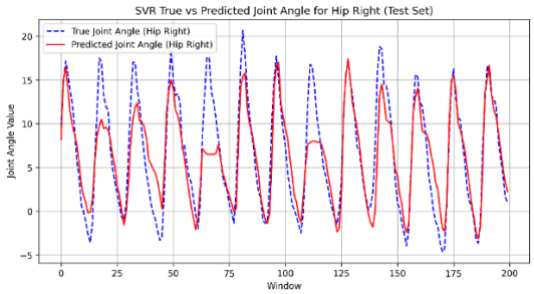

Figure 57: SVR Hip Right (Test Set) adding N kin samples

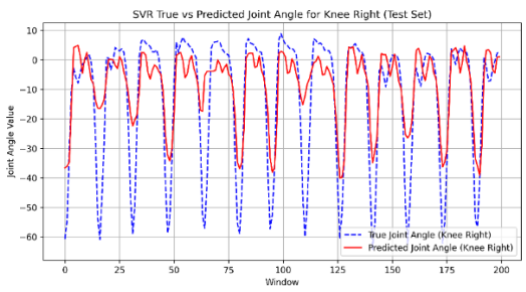

Figure 58: SVR Knee Right (Test Set) adding N kin samples

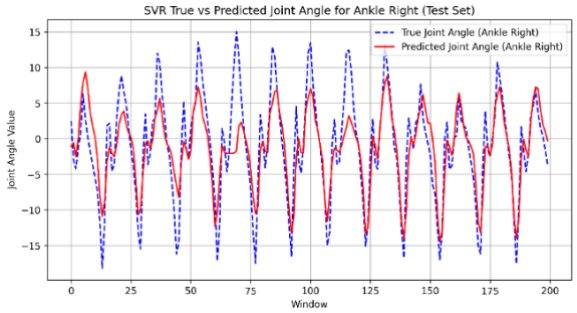

Figure 59: SVR Ankle Right (Test Set) adding N kin samples

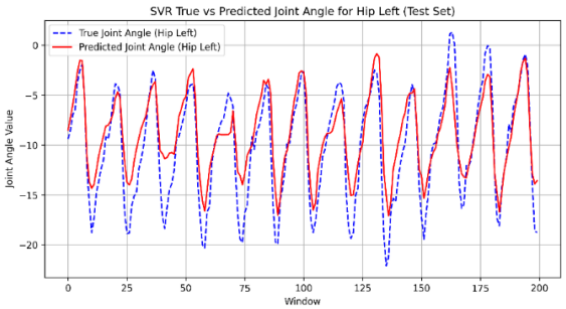

Figure 60: SVR Hip Left (Test Set) adding N kin samples

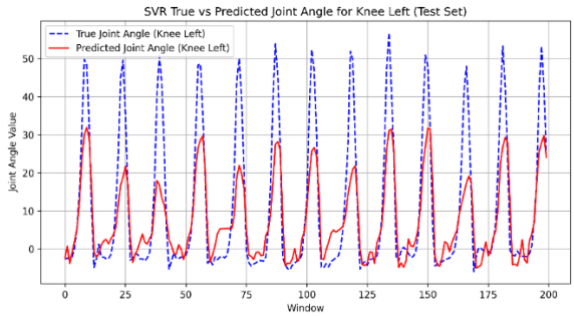

Figure 61: SVR Knee Left (Test Set) adding N kin samples

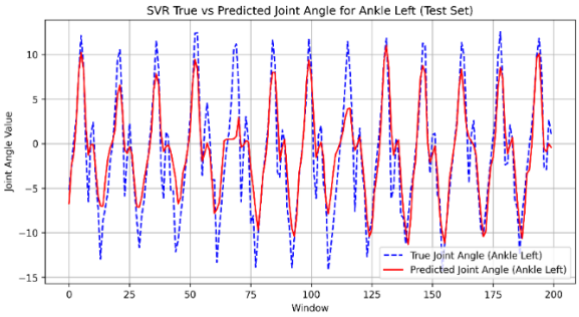

Figure 62: SVR Ankle Left (Test Set) adding N kin samples

Now I present the results obtained from applying the Random Forest model (Figure 63, Figure 64, Figure 65, Figure 66, Figure 67, Figure 68).

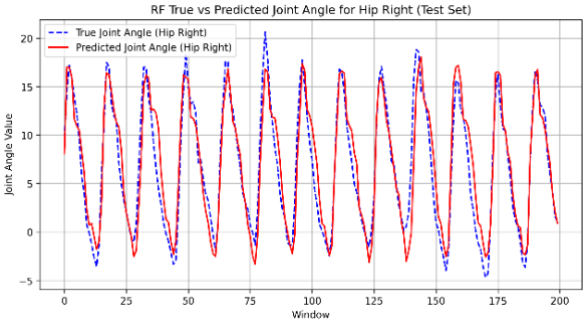

Figure 63: RF Hip Right (Test Set) adding N kin samples

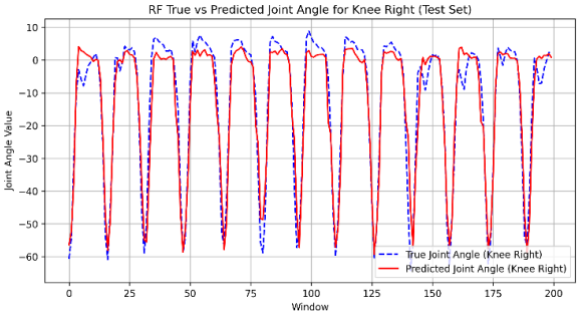

Figure 64: RF Knee Right (Test Set) adding N kin samples

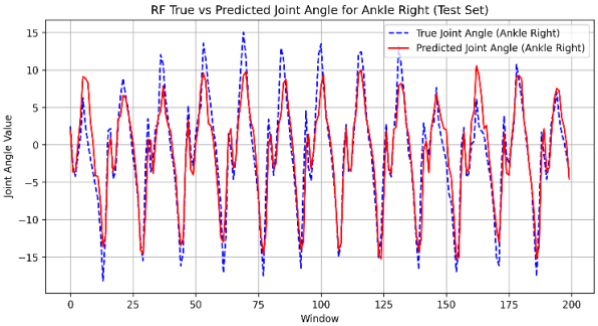

Figure 65: RF Ankle Right (Test Set) adding N kin samples

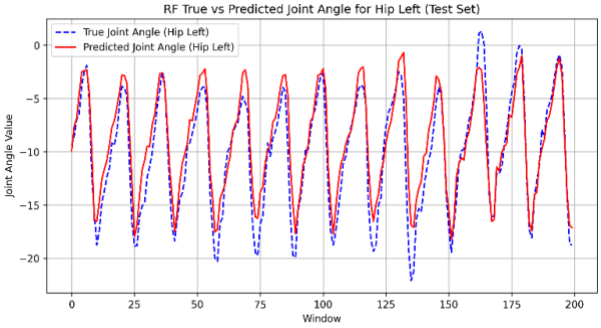

Figure 66: RF Hip Left (Test Set) adding N kin samples

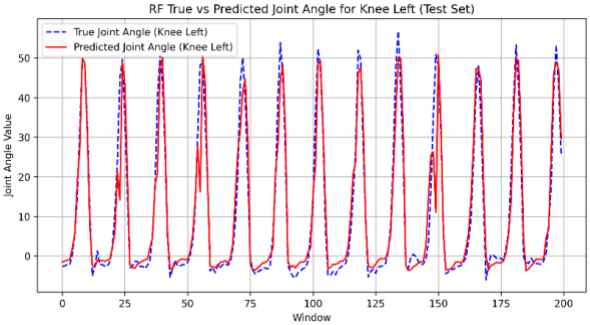

Figure 67: RF Knee Left (Test Set) adding N kin samples

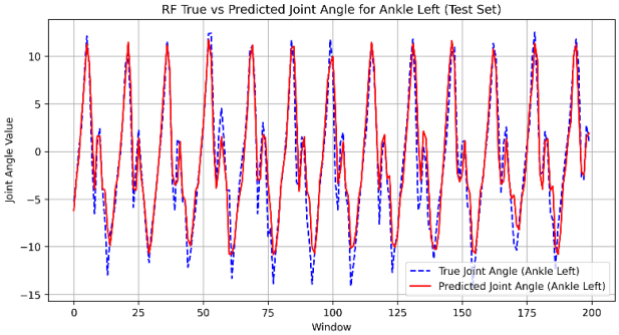

Figure 68: RF Ankle Left (Test Set) adding N kin samples

Following the results obtained from applying the KNN model are shown (Figure 69, Figure 70, Figure 71, Figure 72, Figure 73, Figure 74).

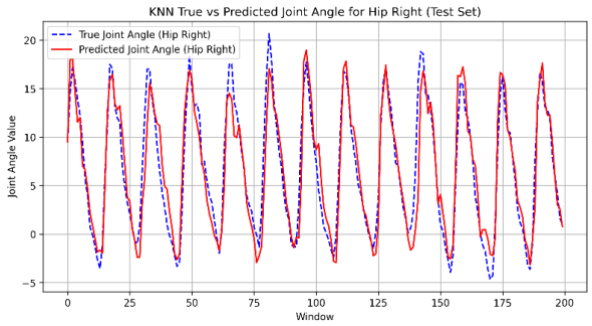

Figure 69: KNN Hip Right (Test Set) adding N kin samples

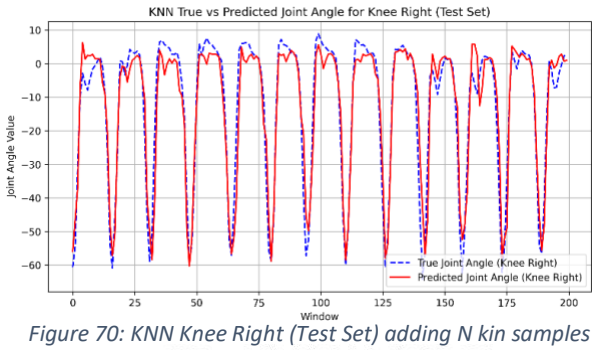

Figure 70: KNN Knee Right (Test Set) adding N kin samples

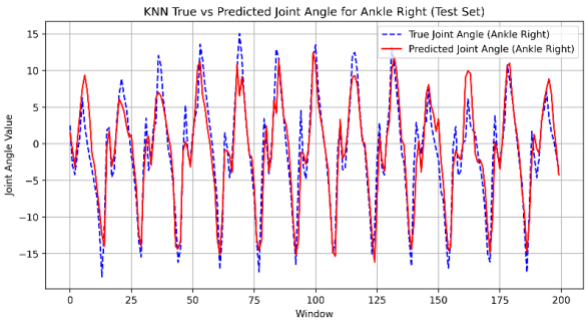

Figure 71: KNN Ankle Right (Test Set) adding N kin samples

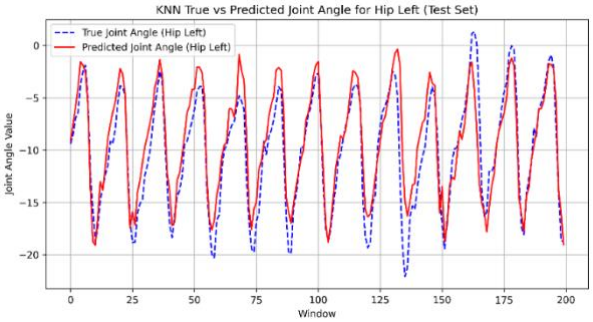

Figure 72: KNN Hip Left (Test Set) adding N kin samples

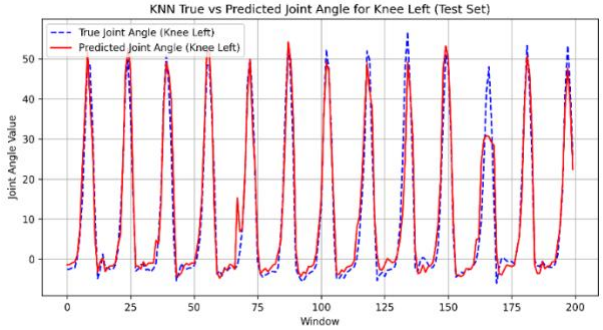

Figure 73: KNN Knee Left (Test Set) adding N kin samples

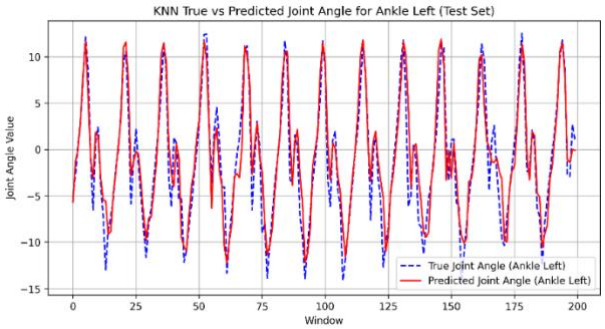

Figure 74: KNN Ankle Left (Test Set) adding N kin samples

Lastly, I show the results obtained from applying the Gradient Boosting regression model (Figure 75, Figure 76, Figure 77, Figure 78, Figure 79, Figure 80).

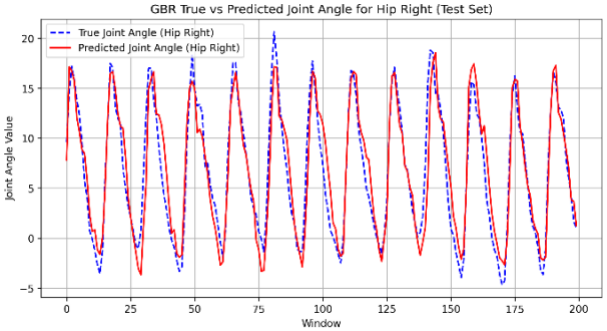

Figure 75: GBR Hip Right (Test Set) adding N kin samples

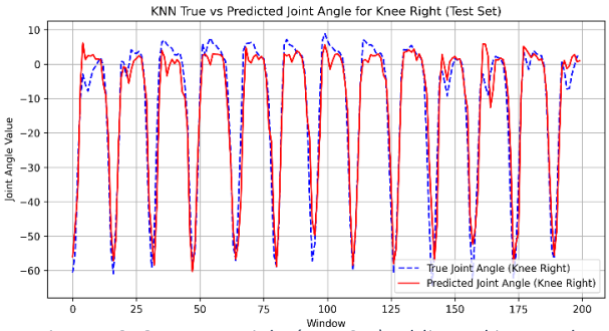

Figure 76: GBR Knee Right (Test Set) adding N kin samples

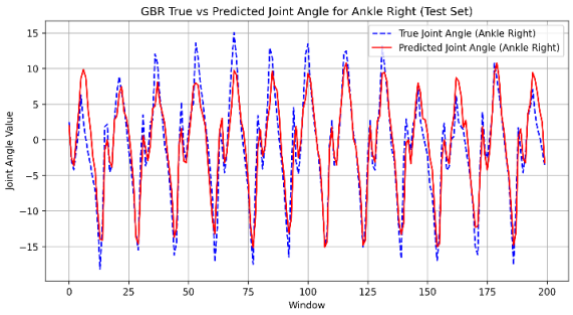

Figure 77: GBR Ankle Right (Test Set) adding N kin samples

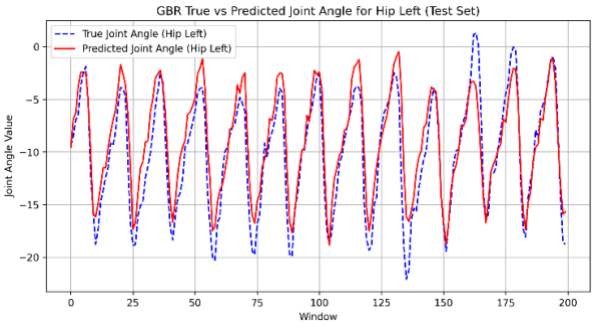

Figure 78: GBR Hip Left (Test Set) adding N kin samples

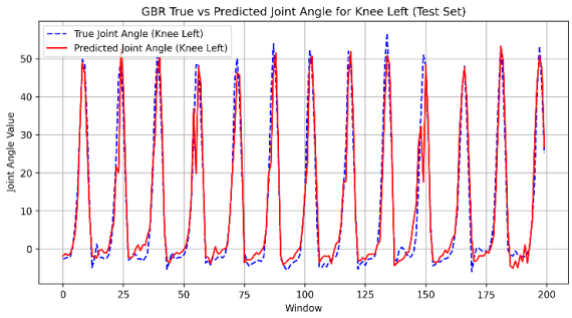

Figure 79: GBR Knee Left (Test Set) adding N kin samples

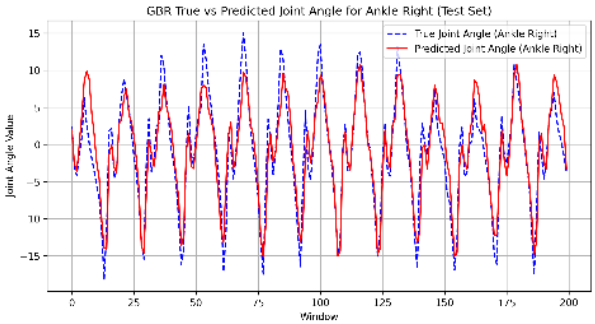

Figure 80: GBR Ankle Left (Test Set) adding N kin samples

From a qualitative point of view, it can be seen that adding the first N=4 samples in each window makes the predictions better. Upon examining the error metrics, the results observed in the graphs are confirmed by the numbers. Specifically, there is a clear reduction in error compared to the case where prediction was made using only EEG data. It is also evident that the knee joint consistently shows the highest error values, in contrast to the hip and ankle joints (Figure 81).

Model	Joint	MSE	MAE	R^2
SVR	Joint 1	9.06986	2.21912	0.78967
SVR	Joint 2	183.24959	9.37401	0.61321
SVR	Joint 3	15.52074	3.03354	0.71052
SVR	Joint 4	7.38348	2.15189	0.75735
SVR	Joint 5	114.48199	7.02055	0.67689
SVR	Joint 6	9.16850	2.29199	0.79627
Random Forest	Joint 1	4.96375	1.72969	0.88489
Random Forest	Joint 2	29.45286	4.02320	0.93783
Random Forest	Joint 3	8.23690	2.22109	0.84637
Random Forest	Joint 4	4.28962	1.65029	0.85903
Random Forest	Joint 5	34.09774	3.29849	0.90376
Random Forest	Joint 6	4.62154	1.57337	0.89731
KNN	Joint 1	4.94035	1.64259	0.88543
KNN	Joint 2	37.68273	4.41711	0.92046
KNN	Joint 3	8.62201	2.26233	0.83919
KNN	Joint 4	5.48043	1.75837	0.81989
KNN	Joint 5	20.98636	2.96622	0.94077
KNN	Joint 6	6.15028	1.82286	0.86334
Gradient Boosting	Joint 1	5.13915	1.74945	0.88082
Gradient Boosting	Joint 2	34.71319	4.32719	0.92673
Gradient Boosting	Joint 3	7.98149	2.20997	0.85114
Gradient Boosting	Joint 4	5.51300	1.84329	0.81882
Gradient Boosting	Joint 5	27.97230	3.17207	0.92105
Gradient Boosting	Joint 6	4.62278	1.64144	0.89728

Figure 81: Models performances

5.3 Stance vs swing predictions

Below I show the graph depicting the division of gait into swing and stance phases, specifically for the right hip joint (Figure 82).

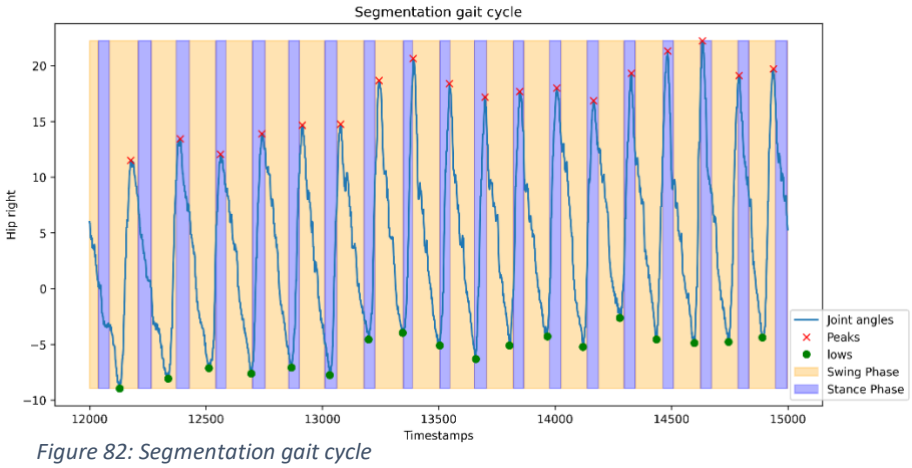

Figure 82: Segmentation gait cycle

Following, the confusion matrices and ROC curve of stance and swing phase prediction using three different classification models (SVR, RF, KNN) (Figure 83, Figure 84, Figure 85, Figure 86, Figure 87, Figure 88).

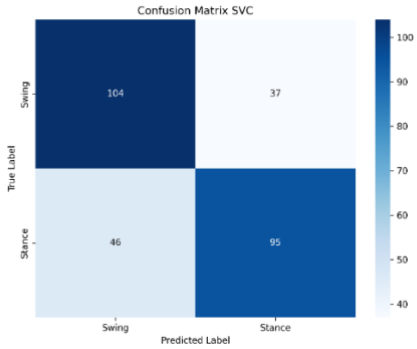

Figure 83: Confusion Matrix SVC

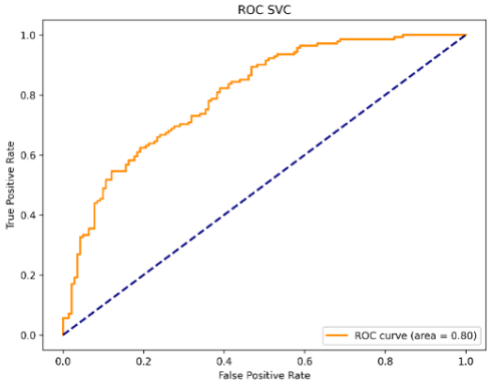

Figure 84: ROC SVC

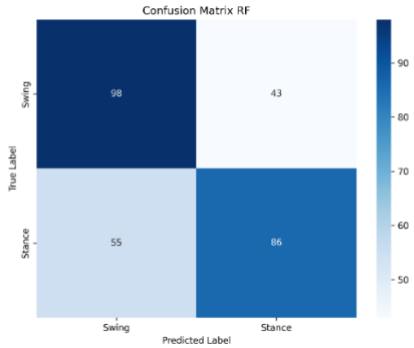

Figure 85: Confusion Matrix RF

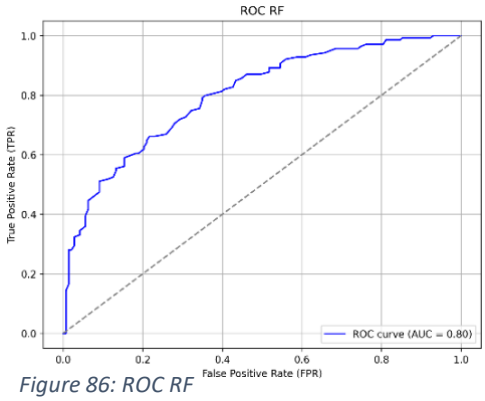

Figure 86: ROC RF

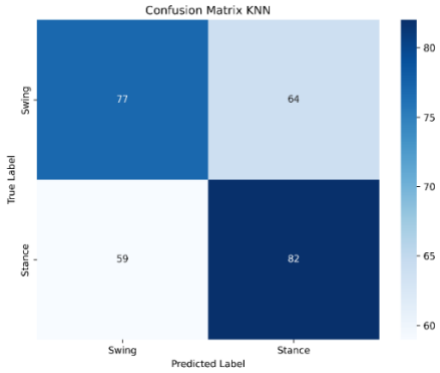

Figure 87: Confusion Matrix KNN

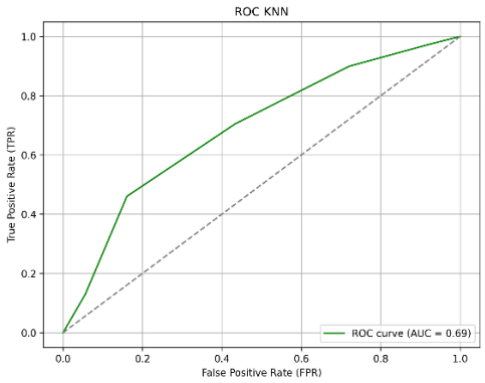

Figure 88: ROC KNN

After analyzing the confusion matrices and ROC curves, it is evident that the Support Vector Classifier (SVC) outperforms the other models in classifying the stance and swing phases using EEG data, achieving an accuracy of 70.57% and an ROC area of 0.80. Random Forest (RF) has an accuracy of 65.25% and an ROC area of 0.75. It demonstrates robustness to noise and serves as a competitive alternative. In contrast, the k-Nearest Neighbors (KNN) model delivers the poorest performance, with an accuracy of 56.38% and an ROC area of 0.62.

5.4 N classes division

Below, I present the results after dividing the angular values into different classes with a certain tolerance and attempting to predict the class to which the last angular sample of each window belonged. Specifically, we discretized the angles into n classes, varying n each time (Figure 88, Figure 89, Figure 90, Figure 91, Figure 92).

The rationale behind discretizing a continuous output and transforming a regression problem into a classification task lies in the fact that achieving extremely precise angle predictions is unnecessary for our purposes. Instead, it suffices to ensure that the prediction error remains within an acceptable margin of error (denoted as Delta Δ, though its exact value has not been defined at this stage). This approach allows us to focus on capturing the general direction or range of the angular values rather than pinpointing their exact value.

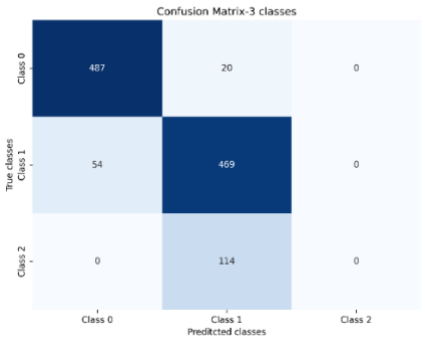

Figure 89: RF Confusion Matrix 3-classes

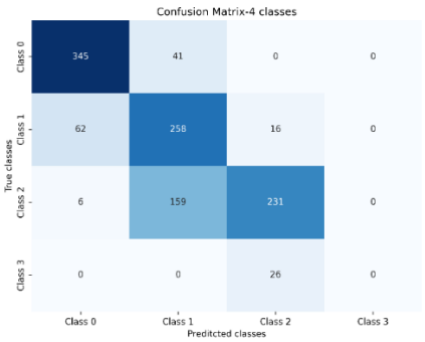

Figure 90: RF Confusion Matrix 4-classes

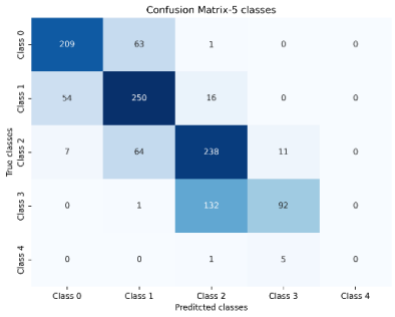

Figure 91: RF Confusion Matrix 5-classes

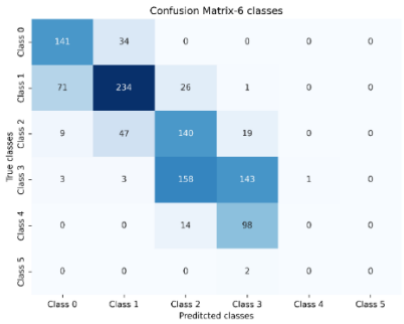

Figure 92: RF Confusion Matrix 6-classes

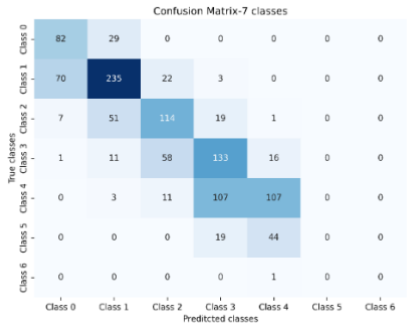

Figure 93: RF Confusion Matrix 7-classes

For each class configuration, we evaluated the accuracy of the Random Forest (RF) model. The results showed a clear trend of decreasing accuracy as the number of classes increased:

- For n=3, the model achieved an accuracy of 84.56%.
- For n=4, the accuracy dropped to 72.90%.
- For n=5, the accuracy further decreased to 68.97%.
- For n=6, the accuracy fell to 57.52%.

- Finally, for n=7, the accuracy was 58.65%.

Below, I show the 3D plot, representing the Mean Square Error of RF model in predicting the last angular sample of each window, as function of the number of kinematic samples considered at the beginning of each window and window size (Figure 94).

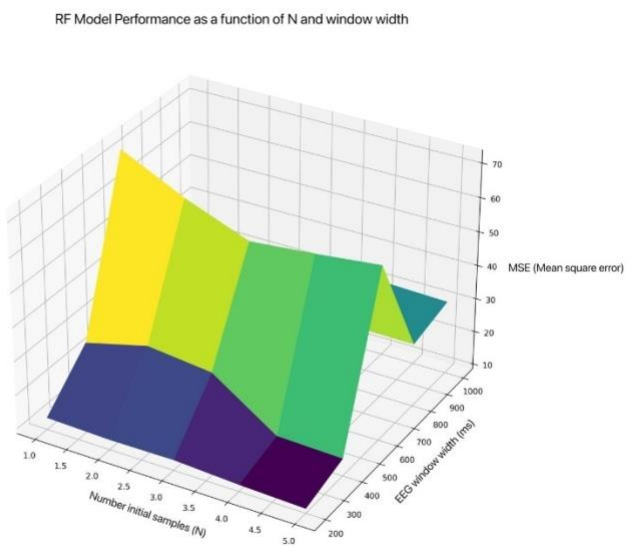

Figure 94: RF Model Performance

It can be observed that for low values of N and intermediate EEG window durations (600-800 ms), the error is very high. As N increases, the MSE progressively decreases, with better performance for high values of N. Short EEG windows (200-400 ms) or very long ones (1000 ms) tend to reduce the MSE, but only when N is sufficiently high.

6 Discussions

As it can be seen from first plots, EEG signals primarily capture the intention or central control of movement rather than directly encoding biomechanical parameters like forces or joint angles. While traditional machine learning algorithms such as SVM, Random Forest, or regression models can identify general patterns—such as distinguishing gait phases or classifying locomotor states—the weak correlation between EEG and specific biomechanical variables, combined with the susceptibility of EEG signals to artifacts and the need for large, high-quality datasets, significantly limits predictive accuracy. Therefore, without the integration of additional sensors (e.g., EMG or inertial measurement units) or leveraging more advanced methods like deep learning, EEG alone is inadequate for precisely predicting complex gait kinematics. SVR (Support Vector Regression) outperforms all other models in predicting joint angles across the board. This is primarily due to its ability to handle non-linear relationships in the data, especially when using the RBF (Radial Basis Function) kernel. The RBF kernel maps the data into a higher-dimensional space, where complex relationships become more separable and can be modeled more effectively. When the relationship between EEG signals and joint angles is non-linear, SVR with an appropriate kernel significantly enhances prediction accuracy. Including the first 4 angular samples in machine learning models improves performance by providing important temporal context that helps predict subsequent values more accurately. These initial samples offer a clear starting point, enabling the model to better understand the movement dynamics and learn the patterns of the gait cycle. As a result, predictions become more stable and precise, as the models benefit from a more coherent and informative sequence of data. However, in this case, Support Vector Regression (SVR) performs slightly worse than the other models. This discrepancy in performance can be attributed to the nature of the models used. While SVR with an RBF kernel is known for its ability to model nonlinear relationships, in this case, decision tree-based models such as Random Forest and Gradient Boosting perform better. Although SVR can capture nonlinearity, it is particularly sensitive to noise and the variability of the data, such as the EEG signals, which are prone to artifacts and physiological fluctuations. In contrast, decision tree models, due to their ability to handle noise and capture nonlinear interactions, adapted better to the complex data, improving generalization. Specifically, the use of ensemble decision trees enabled these models to be more robust and accurate in predictions, outperforming SVR in this context. Moreover, the high R^2 value is a positive indicator, as it suggests that the models explains a significant portion of the variability in the data. In other words, an R^2 value close to 1 means the predictions closely match the observed values, demonstrating the model's strong ability to capture

the relationships between the variables. This reflects their solid performance in accurately representing the data and highlights effectiveness in analysis and prediction. SVC is once again the best model for classification between swing and stance phase. Its performance is largely attributed to the RBF kernel, which enables SVC to effectively model nonlinear relationships and separate classes in the high-dimensional space characteristic of EEG signals. RF falls short compared to SVC in optimizing class separability. Bad performances of KNN are likely due to its simplicity and reliance on distance metrics, which make it less effective in handling the complexity and noise present in EEG data. In conclusion, SVC emerges as the most suitable model for this task, thanks to its ability to manage the complexity and nonlinearity inherent in EEG signals. This highlights the importance of selecting a model that aligns with the specific challenges of the problem at hand. Building on this analysis of model performance, it is crucial to delve deeper into the interaction between the number of initial kinematic samples (N) and the EEG window duration, as illustrated in the 3D MSE plot of the Random Forest model. The plot reveals that higher N values significantly enhance the model's accuracy, as these samples provide direct and reliable information for predicting the final kinematic state. However, the effect of EEG window duration is more nuanced: shorter windows (200-400 ms) and longer ones (1000 ms) tend to reduce the MSE, but only when N is sufficiently large. In contrast, intermediate windows (600-800 ms) result in an increase in MSE, likely due to incomplete temporal context and additional noise. These findings highlight the importance of optimizing both N and EEG window duration to minimize error. The Random Forest model effectively adapts to these interactions due to its ability to handle noise and capture nonlinear dependencies, further demonstrating the versatility of tree-based models in complex and noisy scenarios, such as those involving EEG data. As regards the division of angular values into n different classes, the decline in accuracy with the increasing number of classes can be attributed to the growing complexity of the classification problem. As the number of classes increased, the model faced greater difficulty distinguishing between similar classes, which impaired its generalization ability. Specifically, by reducing the number of classes, the model was able to more clearly separate the categories, thus yielding better performance. On the other hand, with a higher number of classes, the increased granularity of decision boundaries made it more challenging for the model to classify the samples accurately, leading to a reduction in overall accuracy. In summary, this analysis highlighted that the performance of the Random Forest model was strongly influenced by the choice of the number of classes. Fewer classes led to more satisfactory results due to better separability between categories, while an increase in the number of classes made it more difficult for the model to accurately distinguish between samples, resulting in a decrease in accuracy.

7 Conclusions

In the course of this research, we explored the use of EEG to predict gait kinematics in healthy subjects. The results were promising, particularly due to the inclusion of certain kinematic samples at the beginning of each time window, which significantly enhanced the predictive capacity of the model. Notably, the recognition of the stance and swing phases—critical for understanding human gait—yielded satisfactory results, supported by the application of various machine learning methods. However, it is crucial to examine the limitations of these approaches, the challenges encountered during the study, and potential future directions to further refine the system. One major difficulty identified was the quality of the EEG signal, which is often affected by noise. EEG signals are highly susceptible to artifacts from various sources, such as muscle movements or physiological changes, which can negatively impact data quality. Despite the use of advanced filtering and preprocessing techniques, noise remains a significant obstacle. This might explain some discrepancies between the predicted and observed data, particularly during the transition periods between gait phases, where EEG signals tend to be less distinctive. Furthermore, while the classifiers used (e.g., SVM, Random Forest) demonstrated good overall performance, they exhibited certain limitations. For instance, these methods can suffer from overfitting when trained on excessively noisy signals, potentially leading to a loss of generalization in real-world scenarios, where inter-subject variability is higher than in the training samples. Additionally, the temporal resolution and transient nature of EEG signals do not always allow for a clear separation between different gait phases, making predictive accuracy challenging in certain cases. Another important consideration is that, although the results are promising, predicting gait kinematics remains a complex task. The inclusion of kinematic samples at the beginning of each window improved the model's performance but may not suffice to guarantee robust predictions in real-world or clinical contexts. The model's adaptability to physiological and behavioral variations across subjects remains an open question. To delve deeper into the potential of EEG in the context of gait analysis, we implemented a complex experiment combining robotic assistance in walking with a multimodal framework. Integrating robotics and other sensors into a unified system introduced a new dimension to our research, allowing for the investigation of the neural correlates of gait in a more dynamic and realistic setting. However, this complexity also introduced new challenges, particularly in synchronizing signals from multiple sources (EEG, kinematic sensors, robotics). Managing and integrating such multimodal data represents an additional technical hurdle, yet it may pave the way for a more comprehensive understanding of the neural and motor dynamics involved in gait. Looking ahead, there are several avenues for advancing this

research. First, improving EEG signal quality and reducing noise could be a critical step. Advanced preprocessing techniques, such as spatial filters or deep learning-based denoising methods, could enhance signal reliability. Additionally, real-time monitoring systems incorporating advanced sensors could contribute to more accurate gait phase predictions. Employing more sophisticated machine learning models, such as deep learning architectures, including recurrent neural networks (RNNs), could enable better modeling of the temporal and spatial variability in the data. The ability to "learn" the temporal sequence of the data and capture interactions between EEG signals and movements could further improve gait prediction accuracy. Another promising direction is the expansion of the multimodal approach. Integrating EEG signals with data from motion sensors, computer vision systems, and robotics could create a more personalized and effective gait assistance system. This approach would be particularly valuable in rehabilitative settings, where personalized treatment is critical for optimizing outcomes. Finally, applying this system to clinical populations represents an exciting and impactful challenge. Predicting gait in patients with motor impairments, such as those with neurological disorders or spinal injuries, could greatly benefit from robotic assistance technologies combined with real-time EEG analysis, opening new possibilities for rehabilitation. In conclusion, while the results of this research are promising, significant challenges remain in optimizing the use of EEG as a predictive tool for gait. Technical limitations related to signal noise, individual variability, and multimodal data integration require further development. Nevertheless, the proposed approach marks an important step toward an integrated system that could enhance gait assessment and treatment, with potential applications in both healthy and clinical contexts.

Figures index

References

[1] Alessandro Presacco, Larry W. Forrester, Jose L. Contreras-Vidal. Decoding Intra-Limb and Inter-Limb Kinematics During Treadmill Walking From Scalp Electroencephalographic (EEG) Signals. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* ,2012.

[2] Alessandro Presacco , Ronald Goodman, Larry Forrester, Jose Luis Contreras-Vidal. Neural decoding of treadmill walking from noninvasive electroencephalographic signals. *Journal of Neurophysiology,* 2011.

[3] C.S. Sherrington. Flexion-reflex of the limb, crossed extension-reflex, and reflex stepping and standing. *The Journal of Physiology*,1910.

[4] Thomas Graham Brown. The intrinsic factors in the act of progression in the mammal. *The Royal Society,* 1911.

[5] Stefano Tortora, Stefano Ghidoni, Carmelo Chisari, Silvestro Micera, and Fiorenzo Artoni. Deep learning-based BCI for gait decoding from EEG with LSTM recurrent neural network. *Journal of Neural Engineering*, 2020.

[6] Natalie Richer, J.Cortney Bradford, Daniel P. Ferris. Mobile neuroimaging: What we have learned about the neural control of human walking, with an emphasis on EEG-based research. *Elsevier***,** 2024.

[7] Dasa Gorjan, Klaus Gramann, Kevin De Pauw and Uros Marusic. Removal of movement-induced EEG artifacts: current state of the art and guidelines. *Journal of Neural Engineering,* 2022.

[8] M.L. Shik and G. N. Orlovsky. Neurophysiology of locomotor automatism. Physiological reviews, 1976.

[9] W. Widajewicz, B. Kably, and T. Drew. Motor cortical activity during voluntary gait modifications in the cat. *Journal of neurophysiology*, 1994.

[10] S. Della Sala, A. Francescani, H. Spinnler. Gait apraxia after bilateral supplementary motor area lesion. *Journal of Neurology, Neurosurgery and Psychiatry,* 2002.

[11] S. Rossignol, R. Dubuc and J.P. Gossard. Dynamic Sensorimotor Interaction in Locomotion. *Physiological reviews,* 2006.

[12] T. Bulea, A. Kilicarslan, R. Ozdemir, W. H. Paloski, J.L. Contreras-Vidal. Simultaneous Scalp Electroencephalography (EEG), Electromyography (EMG), and Whole.body Segmental Intertial Recording for Multi-modal Neural Decoding. *National Library of Medicine,* 2013.

[13] B. Calancie, B. Needham-Shropshire, P. Jacobs, K. Willer, G. Zych, B. A. Green. Involuntary stepping after chronic spinal cord injury: Evidence for a central rhythm generator for locomotion in man. *Brain, Volume 117, Issue 5*, 1994.

[14] M.R.Dimitrijevic, Y. Gerasimenko, M. M. Pinter. Evidence for a Spinal Central Pattern Generator in Humans. *The New York Academy of Sciences,* 2006.

[15] N. Kawashima, D. Nozaki, M. O. Abe, and K. Nakazawa. Shaping Appropriate Locomotive Motor Output Through Interlimb Neural Pathway Within Spinal Cord in Humans. *Journal of Neurophysiology,* 2008.

[16] P. L. Sheridan, J. M. Hausdorff. The Role of Higher-Level Cognitive Function in Gait: Executive Dysfunction Contributes to Fall Risk in Alzheimer's Disease. *Dementia and Geriatric Cognitive Disorders, Volume 24, Issue 2,* 2007.

[17] G. Yogev-Seligmann, J. M. Hausdorff, N. Giladi. The role of executive function and attention in gait. *Official Journal of the International Parkinson and Movement Disorder Society,* 2007.

[18] K. Gramann, J. T. Gwin, D. P. Ferris, K. Oie, Tzyy-Ping Jung, Chin-Teng Lin, Lun- De Liao and S. Makeig. Cognition in action: imaging brain/body dynamics in mobile humans. *Reviews in the Neurosciences*, 2011.

[19] B. J. Thio, W. M. Grill. Relative contributions of different neural sources to the EEG. *Elsevier,* 2023.

[20] M. Teplan. Fundamentals of EEG measurement. *Measurement Science Review, Volume 2, Section 2,* 2002.

[21] G. Pfurtscheller, F.H. Lopes da Silva. Event-related EEG/MEG synchronization and desynchronization: basic principles. *Elsevier*, 1999.

[22] S. Makeig. Auditory event-related dynamics of the EEG spectrum and effects of exposure to tones. *Elsevier,* 1993.

[23] T. H. Petersen, M. Willerslev-Olsen, B. A. Conway, J. B. Nielsen. The motor cortex drives the muscles during walking in human subjects. *The Journal of Physiology, Volume 590, Issue 10,* 2012.

[24] A. T. Winslow, J. Brantkey, F. Zhu, J. L. Contreras Vidal, H. Huang. Corticomuscular coherence variation throughout the gait cycle during overground waking and ramp ascent: A preliminary investigation. *IEEE Xplore,* 2016.

[25] C. Park, J. Seo, D. Kim, B. Abibullaev, H. Kwon, Y. Lee, M. Kim, K. An, K. Kim, J. S. Kim, E. Y. Joo, S. B. Hong. EEG source Imaging in Partial Epilepsy in Comparison with Presurgical Evaluation and Magnetoencephalography. *Journal of Clinical Neurology,* 2015.

[26] Y. Zhang, X. Ji, S. Zhang. An approach to EEG-based emotion recognition using combined feature extraction method. *Elsevier,* 2016.