NATIONAL TECHNICAL UNIVERSITY OF ATHENS
SCHOOL OF ELECTRICAL AND COMPUTER ENGINEERING
SCHOOL OF MECHANICAL ENGINEERING

INTERDISCIPLINARY POSTGRADUATE PROGRAMME
"Translational Engineering in Health and Medicine"

Machine Learning Assessment of EEG Data in a fatigue-related n-back task

Postgraduate Diploma Thesis

Ioannis Charalampous

Supervisor: *George K. Matsopoulos , Professor, NTUA*

Athens, June, 2025

NATIONAL TECHNICAL UNIVERSITY OF ATHENS
SCHOOL OF ELECTRICAL AND COMPUTER ENGINEERING
SCHOOL OF MECHANICAL ENGINEERING

INTERDISCIPLINARY POSTGRADUATE PROGRAMME
"Translational Engineering in Health and Medicine"

Machine Learning Assessment of EEG Data in a fatigue-related n-back task

Postgraduate Diploma Thesis

Ioannis Charalampous

Supervisor: *George K. Matsopoulos , Professor, NTUA*

The postgraduate diploma thesis has been approved by the examination committee on 24/6/2025 (exam day: 24 June 2025)

1st member	*2nd member*	*3rd member*
Prof. G. Matsopoulos	*Prof. P. Tsanakas*	*As. Prof. Ch. Manopoulos*
NTUA	*NTUA*	*NTUA*

Athens, June 2025

Ioannis Charalampous
Graduate of the Interdisciplinary Postgraduate Programme,
"Translational Engineering in Health and Medicine",
Master of Science,
School of Electrical and Computer Engineering,
National Technical University of Athens

to my family

Abstract

Mental fatigue significantly seriously impacts cognitive functions and decision-making, especially in safety-related environments like aviation, medicine, and transportation. This thesis presents a methodology for detecting mental fatigue from electroencephalography (EEG) signals recorded while doing an n-back working memory task. A hybrid deep learning model that combines a convolutional neural network (CNN) and bidirectional long short-term memory network (BiLSTM) was developed to classify EEG signals as "fatigued" and "rested" states. In order to enhance the spatial resolution of the EEG signals, source localization was performed using the sLORETA algorithm, which projects scalp-recorded activity onto cortical surfaces. The model was trained and tested using 10-fold cross-validation, and achieved an average accuracy of 91.55%, demonstrating that it is robust and generalized across subjects. Explainability was also ensured through SHapley Additive exPlanations (SHAP), which provided insight regarding the most salient cortical sources that are responsible for the model predictions. The analysis highlighted contributions from frontal and parietal regions, that are consistent with neuroscientific findings on fatigue-related changes in brain activity. The dataset was collected from recordings of the participants in both rested and sleep-deprived conditions, enabling the model to learn discriminative patterns associated with mental fatigue. This work not only offers a high-performing and also explainable model for EEG-based fatigue assessment but also reduces the gap between deep learning and neuroscience by connecting machine learning predictions with physiologically meaningful brain processes.

KeyWords

Mental fatigue, EEG, n-back task, deep learning, CNN-BiLSTM, source localization, sLORETA, SHAP, cognitive workload, model explainability

Acknowledgements

I would first like to thank professor George K. Matsopoulos as well as the postdoctoral researcher Ioannis Kakkos for the supervision of this thesis, for the opportunity they gave me to work on this thesis and for their guidance and the excellent cooperation we had. Finally, I would like to thank my family and friends for their guidance and moral support throughout the years.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Introductory Note

Mental fatigue is known to have a significant impact on the cognitive performance, decision making, and productivity daily on humans across various domains. With prolonged mental exertion, mental fatigue is inevitable in modern lifestyle, manifesting as reduced efficiency in task execution, slower reaction times, impaierd judgment, and an increased likelihood of errors [7]. Given its profound consequences, particularly in high-stakes environments such as aviation, healthcare, and industrial operations, understanding and assessing mental fatigue have become key areas of research interest. One important part of this research involves electroencephalography (EEG), a non-invasive neuroimaging technique, which is widely used in neuroscience for studying mental states like fatigue. EEG can provide real-time view of neural activity with high temporal resolution that can capture patterns important to correlate cognitive workload and fatigue.

Based on EEG research mental fatigue is often associated with changes in brain oscillatory activity, particularly in alpha and theta frequency bands. These oscillations are important marks of cognitive load that provides a physiological basis for classifying mental states [8]. Despite its many advantages EEG signals are complex and contain noise, necessitating advanced preprocessing and modeling techniques to extract meaningful features, Cognitive workload, closely linked to mental fatigue, is commonly manipulated through working memory tasks such as the n-Back task. This task involves recalling and processing stimuli presented N steps earlier in a sequence and is extensively used in cognitive neuroscience to evaluate working memory capacity [9]. The n-Back task is particularly suitable for fatigue assessment as it allows for controlled manipulation of cognitive load, making it a good framework for studying changes in neural activity that are related to fatique.

1.2 Purpose of the Diploma Thesis

The primary objective of this thesis is to develop a framework for assessing mental fatigue using EEG signals recorded during an n-Back task. The study introduces a deep learning model based on a hybrid architecture combining Convolutional Neural Networks (CNNs) and Bidirectional Long Short-Term Memory (BiLSTM) networks. The CNN component captures spatio-temporal patterns across cortical sources, while the BiLSTM component models temporal dependencies, in order to detect sequential neural dynamics that are associated with fatigue. To enhance the relevance of our EEG features, source localization preprocessing was employed with the help of sLoretta program. This method projects EEG signals that are acquired from the subjects scalp onto cortical sources, with the goal to improve spatial resolution and isolate neural activity linked to cognitive processes involved in the n-Back task.

Explainability was introduced in this work woth the incorporation of SHapley Additive exPlanations (SHAP) to ensure interpretability. SHAP can atribute the model predictions to specific EEG features,giving us important insights into the neural mechanisms of mental fatigue.WIth this explainability the trustworthiness of the model is enhanced with the goal to connect the machine learning predictions with neuroscience results. The dataset that was collected that helped achieving this goal includes EEG recordings from participants under the experimental conditions of rested (prior to sleep deprivation) and fatigued (after 24 hours of sleep deprivation). By employing advanced preprocessing, hybrid deep learning techniques, and explainability tools, this thesis aims to achieve high classification accuracy while ensuring the model's predictions are physiologically meaningful.

1.3 Contents of the Diploma Thesis

This thesis is organized into seven chapters, each addressing a critical aspect of the study on mental fatigue assessment using EEG signals. The structure ensures a logical flow, from foundational knowledge to experimental results and conclusions.

1. Introduction: The introduction outlines the significance of mental fatigue in cognitive performance and its broader implications. It presents the scope and objectives of the study,and concludes with an overview of the thesis structure.

2. Theoretical background: This chapter is the foundation for the study. It shows brain anatomy, focusing on the lobes and frequencies relevant to cognitive workload and fatigue. It also introduces the concept of mental fatigue, its physiological basis, and its impact on neural activity. The role of EEG in neuroscience and the cognitive demands of working memory tasks, particularly the n-Back task, are also explained.

3. Related work: In this chapter previous research on EEG-based mental fatigue assessment is reviewed. It highlights the use of application of machine learning models in classifying mental states, discussing their advantages and limitations. The chapter

shows the foundation of applied methodology while also identifies research gaps that this thesis addresses.

4. Methodology: The methodology chapter describes the dataset, including experimental conditions and inclusion criteria. It details the preprocessing pipeline, focusing on source localization using sLoretta, segmentation, and normalization. The proposed CNN-BiLSTM model is presented, along with its architecture, hyperparameters, and the SHAP-based explainability framework.

5. Experimental Procedure and Results: This chapter outlines the experimental setup, including 10-fold cross-validation for robust model evaluation. It presents the results of the classification performance (rested vs. fatigued) using metrics such as accuracy and confusion matrices. SHAP analysis is discussed in detail, highlighting the most influential EEG features for the model's predictions.

6. Conclusions and Future Directions: The concluding chapter summarizes the study's findings, emphasizing the model's performance and interpretability. It discusses limitations, such as dataset size and individual variability, and proposes future research directions, including real-time fatigue detection and expanding the approach to other cognitive tasks.

Chapter 2

Theoretical background

This chapter lays the theoretical foundations for the complex and multidimensional field of the research topic of detecting and classifying mental fatigue using electroencephalography (EEG) data. Understanding the complex functioning of the human brain and the capabilities of cognitive processing is based on many years of research in neuroscience, psychology and, more recently, computational modelling. This work not only involves interpreting complex physiological signals, but also requires understanding the theoretical foundations of the brain and its functions.

2.1 Mental fatigue

Mental fatigue is defined as temporary decrease in cognitive effectiveness which occurs when a person engages in lengthy mentally demanding activities. Mental fatigue is different from physical fatigue because physical fatigue is primarily the product of muscle usage but mental fatigue results from the prolonged use of neural systems which are responsible for maintaining attention and working memory and decision-making [10]. Such deficits are usually manifested in decreased response time, reduced motivation, and both increased error rate and decreased capacity for sustained concentration over long periods. Such impairments are especially bad in complicated real world environment, such as controlling air traffic, nursing a patient or operating heavy machinery, because one has to remain alert to guarantee efficiency and safety. At the neural level, mental fatigue is most commonly linked with changes in brain's organization of activity between cortical areas that function dynamically. The frontal cortex which is central to executive function will show decreased activation in fatigue, which corresponds to the decline in cognitive control. Further, parietal areas that are connected with the processing of information and the maintenance of focus may also become less active with time during tasks that are repetitive or time-long. Using these networks repeatedly may lead to exhaustion of the brain's capacity to effectively manage its resources, which in turn leads to suboptimal performance [11].

Electroencephalography (EEG) offers a non-invasive glimpse of such cortical develop-

ments. By measuring electrical activity at many scalp sites, researchers can measure each moment of fluctuation in neural activity that reflects progression from an alert to a more fatigued state. Although some research breaks down EEG data into frequency bands, several newer approaches consider raw, time-domain signals. Source localization techniques, such as sLORETA that this thesis utilizes, can further refine which cortical generators, like the frontal or parietal regions that are most strongly affected, adding spatial detail to the temporal information in the EEG recordings. In experimental settings, mental fatigue is frequently induced and measured using continuous performance tasks like the n-back working memory paradigm. As participants progress through repeated trials, they often display lengthened reaction times, elevated error rates, and subjective sensations of strain. These markers collectively indicate a reduced capacity to juggle or update information in working memory. However, the degree and onset of such performance decrements can vary considerably among individuals, influenced by factors such as prior rest, personal resilience, and stress levels.

Recognizing these individual differences has spurred interest in personalized and data-driven modeling of fatigue states. Deep learning models that work directly with raw EEG time-series data rather than relying solely on hand-crafted features that offer a promising route to detect emerging fatigue in real time. Identifying subtle neural patterns well before marked behavioral lapses occur has practical benefits such as implementing brief rest breaks or rotating tasks can mitigate risks in professional scenarios that demand continuous alertness. In this way, real-time EEG-based monitoring of mental fatigue contributes not only to theoretical insights into cognitive resource allocation but also to tangible, preventative strategies in domains where sustaining mental performance is paramount.

2.2 Structure of the brain

The human brain, widely regarded as the most complex organ in the body, is the center of our cognitive capacities. This section serves as an explanation of the complex structure of the brain, providing an essential basis for understanding its multifaceted functionality, particularly in the context of cognitive processing and mental fatigue. Understanding the structure of the brain allows us to better appreciate the genesis of the electrical activities recorded by the EEG and thus the very foundation on which EEG-based mental fatigue is based. This overview of brain structure, especially the macro-level organization into lobes and frequencies, provides the biological and neurological background necessary for understanding mental fatigue.

2.3 Brain Anatomy

The human brain is a complex and dynamic organ, consisting of interconnected regions that collectively enable cognitive processes, including attention, working memory,

and decision-making. Understanding the anatomical basis of these processes is crucial for examining how cognitive workload is managed and how mental fatigue disrupts these functions. This section focuses on key brain regions involved in cognitive workload, highlighting their roles and relevance to tasks requiring sustained mental effort. Although we now know that most brain functions rely on many different areas of the whole brain working together, it is still true that each lobe performs most of certain functions. Figure 2.1 below shows a separation of the functions that each area of the brain does.

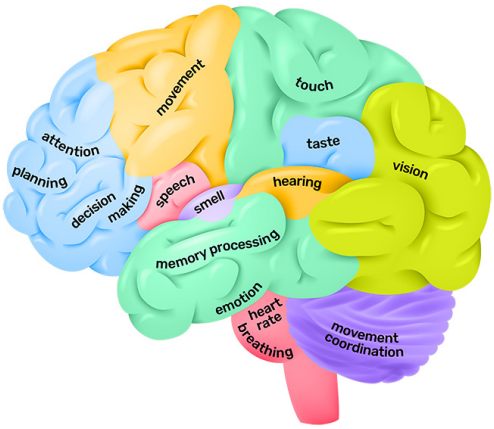


Figure 2.1: Functions of the brain lobes [1]

1. Frontal lobe: the frontal lobe is the largest lobe of the brain, comprising almost a third of the surface area of the hemisphere. It is located largely in the anterior cranial fossa of the skull, resting on the orbital plate of the frontal bone. The frontal lobe is the most anterior part of the cerebral hemisphere and is separated from the parietal lobe posteriorly by the central fissure and from the temporal lobe posteriorly by the lateral fissure (Sylvian fissure).

(a) Region of Frontal Lobe (b) Sylvian fissure

Figure 2.2: Frontal Lobe [2]

2. Parietal lobe: the parietal lobe is located just below the parietal bone, behind the frontal lobe and in front of and above the temporal and occipital lobes. It plays an

important role in integrating sensory information from different parts of the body. It plays an important role in spatial sensation and navigation, object manipulation and number representation.

Figure 2.3: Parietal lobe [3]

Then the inferior parietal lobule (IPL) supports the manipulation of information stored in working memory. This region is particularly active during tasks that require updating or transforming mental representations, such as remembering and comparing sequential stimuli in the n-back test.

2.4 Electroencephalography

Electroencephalography (EEG) is a key tool in the neuroscientific study of brain activity. With its roots dating back to the early 20th century, when German psychiatrist Hans Berger first recorded human brain waves, EEG was first recorded in the animal brain in 1875 by Richard Caton. It captures the fluctuations in voltage generated by neuronal activity, specifically the postsynaptic potentials of pyramidal neurons in the cerebral cortex. This electrical activity is detected by electrodes placed on the scalp and is recorded as continuous waveforms reflecting the dynamics of neural processing. EEG is the most widely used signal acquisition method because of its high temporal resolution, safety and ease of use. EEG has low spatial resolution and is non-stationary in nature.

EEG signals provide insights into brain function with exceptional temporal resolution, often on the millisecond scale. This makes EEG particularly suited for studying fast-occurring cognitive processes such as attention shifts, decision-making, and working memory. EEG signals though are sensitive to artifacts caused by eye blinks, eye movements, heartbeat, muscle activities and power line interference [12].

2.4.1 10-20 system

The international 10-20 system is a method used to standardize the placement of EEG electrodes, paving the way for consistency and reproducibility in EEG studies. The 10-20 system uses anatomical landmarks to standardise the placement of electroencephalography (EEG) electrodes. The system is based on the relationship between the position of the electrodes and the underlying region of the the underlying cerebral cortex, while ensuring that all areas of the brain are covered. The nomenclature ¨10-20' refers to the actual distances between adjacent electrodes, which is either 10% or 20% of the total front-to-back or right-to-left skull distance [13].

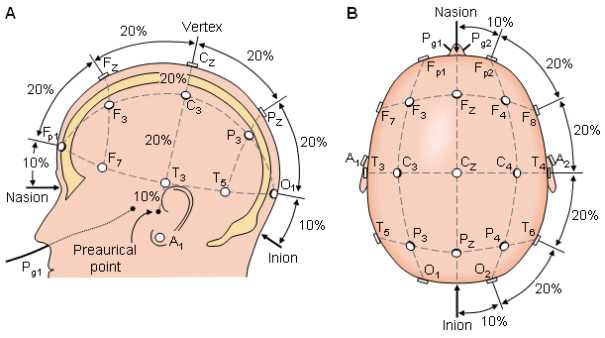


Figure 2.4: Electrode distance in 10-20 system

Chapter 3

Relevant work

Mental workload shows the proportion of limited cognitive resources a person allocates while performing a task. Quantifying mental workload objectively has become really important in domains where safety is critical such as aviation, road transport, air-traffic control because either overload or underload can degrade performance and jeopardise ones safety. Closely allied to mental workload is mental fatigue, the progressive loss of efficiency that accompanies sustained cognitive effort [14]. Mental fatigue manifests as slowed reaction time, lapses of attention and reduced motivation, and neuro-physiologically as drifts in frontal-midline theta and rising delta power during prolonged vigilance [15]. Over the past decade AI has emerged as the leading computational framework for modelling both mental workload and fatigue, largely because modern learning algorithms can capture the complex, non-linear relationships that link neuro-physiological signals to hidden cognitive states [16].

The first EEG-based mental workload systems relied on hand-engineered features such as band power, Hjorth parameters and autoregressive coefficients, which were fed to conventional classifiers including support vector machines or random forests [16]. Although these pipelines achieved reasonable two-class accuracies, they demanded labour-intensive feature design and were highly sensitive to recording noise. Deep learning alleviates many of these limitations by learning hierarchical, task specific representations directly from raw or minimally pre-processed EEG [17]. Convolutional and recurrent networks are able to discover relevant spatio-temporal patterns automatically and routinely outperform classical approaches in multi-class settings. Chakladar and colleagues, for example, combined a bidirectional LSTM with a Grey-Wolf optimiser for feature selection and reached 86% accuracy on a three-level n-back protocol [18]. Fan et al. introduced EEG-TNet which is a 3-D depth-wise separable convolutional network that simultaneously captures frequency, spatial and temporal information and reported subject-dependent accuracies above 99% on binary workload discrimination [19]. More recently Gupta et al. fused model-free functional-connectivity matrices (Mutual Information and Phase-Locking Value) with a lightweight CNN and achieved an average three-class accuracy of about 81%, the best performance yet for subject-specific workload classification [20].

One theoretical contribution of AI to workload research is its capacity for representation learning. Variational auto-encoders followed by spatial-attention modules compress noisy EEG topographies into latent codes where workload-specific boundaries become linearly separable, a strategy that consistently improves downstream CNN-BLSTM classifiers [21]. Hybrid pipelines that integrate robust decomposition (RLMD), meta-heuristic optimisation and ensemble learners can further exploit complementary frequency information, pushing accuracies beyond 97% on two public workload datasets [22].

Because mental workload reflects coordinated activity of distributed neural assemblies, recent work has shifted from single-sensor features to functional-connectivity graphs. Model-free connectivity measures such as mutual information, phase-locking value and phase-transfer entropy provide weighted adjacency matrices that quantify statistical and directed interactions between cortical regions. Convolutional networks operating directly on these matrices (or on graph representations via graph convolution) learn topological descriptors like edge weights, clustering coefficients, network efficiency, which can capture changes in fronto-parietal communication observed under high load [20]. Such graph-based deep models narrow the explanatory gap between EEG and slower neuro-imaging modalities, while maintaining millisecond resolution.

Despite these advances, several challenges remain. Inter-subject variability in EEG and in functional connectivity still limits generalisation domain-adversarial training and meta-learning have been proposed to mitigate this problem, yet large-scale validation is lacking. Data scarcity and class imbalance hamper deep models that thrive on abundant, balanced samples self-supervised pre-training and generative augmentation are promising counter-measures. Real-time deployment requires rapid inference on portable hardware and here the combination of model-free connectivity metrics with shallow CNNs is important because it preserves accuracy while reducing computational load [20]. Finally, interpretable AI is essential for regulatory approval and user trust. Saliency mapping, layer-wise relevance propagation and attention visualisations now reveal that many deep models base their decisions on canonical mental workload correlates such as frontal midline theta and parietal alpha suppression, thereby providing neuro-physiological validity.

The representative papers reviewed below illustrate the trajectory of this field from carefully engineered pipelines rooted in classical power-spectral analysis to sophisticated end-to-end neural architectures augmented with attention mechanisms, evolutionary feature selection or functional-connectivity representations. For example Xu et al. [23] introduced one of the most compact yet effective feature sets for fatigue discrimination by combining relative band power with fuzzy entropy. In a controlled 2-back paradigm, EEG data of partocipants was recorded at three stages (baseline, fatigue induction and recovery). Relative band power captured the redistribution of energy across delta–theta versus alpha–beta ranges that are widely accepted markers of rising cognitive strain while fuzzy entropy quantified the loss of signal complexity as mental resources waned. When these two descriptors were supplied to an Extreme-Gradient-Boosting (XGBoost) classifier, the model achieved 92.4 % average accuracy for these three fatigue levels. Complementing

that single-task approach, Xing et al. [24] tackled cross-task generalisation which is a difficult subject for EEG systems by training on a 2-back dataset and testing on a mental-arithmetic (MA) dataset, and vice-versa. They retained fuzzy entropy as the core feature but adopted a linear-kernel SVM classifier. With only eight participants and 16 EEG channels, the framework still reached a high mean accuracy across the two-task mismatch, demonstrating that entropy-based markers generalise reasonably well when spectral content differs between tasks. Notably, training on the 2-back data and testing on MA yielded slightly higher performance, hinting that the n-Back paradigm elicits more consistent fatigue signatures than arithmetic computation.

Recognising the limitations of manually engineered features and simple classifiers, Zhang et al. [25] proposed a two-stream neural network (TSNN) that learns spectral and temporal EEG patterns in parallel. One branch ingests Welch-derived topographic power maps and the other processes raw event-related segments via a temporal convolutional network (TCN). Fusion of the two streams boosted overall classification to 91.9 % for three workload classes which is substantially higher than the low accuracy of the temporal-only branch. Deconvolution visualisations further revealed that the spectral stream captures theta-alpha power shifts, whereas the temporal stream is sensitive to reductions in P3 and P2 amplitudes, providing interpretable neuro-cognitive evidence for the learned representations.

An important work that inspired this thesis was that of Su et al. [26] where they extended the spatio-temporal idea with a CNN-LSTM hybrid trained on wavelet-denoised EEG from a prolonged 2-back protocol. Their pipeline first extracts spatial patterns through 2D convolutions, then models long-range dynamics with sequential LSTM layers. On a sample of 18 subjects, the network delivered 97.1 % overall accuracy and 97.8 % sensitivity when differentiating awake, mild-fatigue and severe-fatigue states. The authors linked the model's success to its ability to detect frontal-central increases in delta–theta power and concurrent alpha–beta suppression, that are classical markers of fatigue that became pronounced after multiple five-minute task blocks. Another convolutional based architecture which still relies on discrete wavelet preprocessing, was that of Siddhad et al. [27]. They distributed the handcrafted stages entirely by adapting the ConvNeXt vision backbone to 1-s EEG windows. After reducing channel counts and kernel sizes, their model surpassed SVM, EEGNet, TSception and a transformer baseline on the STEW (SIMKAP) dataset, achieving 95.8 % binary and 95.1 % three-class accuracy. Confidence-interval plots showed consistent gains across participants, suggesting that modern image-architecture design principles, including depthwise separable convolutions and large effective receptive fields can translate well to noisy electrophysiological data.

A different unsupervised route was explored by Chakladar et al. [21], who combined a variational auto-encoder (VAE) with a spatial–temporal attention block. The VAE first produces a compact latent representation of each EEG segment and attention then highlights those latent dimensions that are most relevant for workload classification. The VAE and attention pipeline outperformed handcrafted-feature models on two public mental-

arithmetic datasets and maintained interpretability with saliency maps that emphasized frontal and parietal regions that are repeatedly implicated in executive load. Ablation experiments showed that removing either the VAE or the attention module degraded the accuracy, confirming the synergy of deep latent encoding and focused feature weighting. Another solution by Fan et al. [19] presented EEG-TNet, an end-to-end architecture that preserves the raw time dimension while convolving across channels, thereby capturing micro-temporal cues often lost in pooling. Compared with shallower CNNs and LSTMs, EEG-TNet achieved the best fold-wise performance on both subject-specific and cross-subject settings of the n-Back workload benchmark. Saliency visualisation indicated strong reliance on mid-frontal theta bursts and posterior alpha suppression—consistent with cognitive-control literature and underscored the importance of fine-grain temporal kernels that avoid excessive down-sampling.

Hybrid evolutionary deep approaches have also shown promise. For example Das Chakladar et al. [18] applied a grey-wolf pptimiser (GWO) to select an optimal subset from a broad library of spectral, statistical and entropy-based features, which were then passed to a stacked BLSTM-LSTM. The evolutionary filter removed redundant inputs and enhanced interpretability by surfacing theta-alpha ratios and approximate entropy as dominant discriminators. The resulting model delivered 86.3 % accuracy on a three-class SIMKAP workload dataset and proved more compact than full-feature or hand-pruned baselines. Finally, Gupta et al. [20] analyzed workload detection through functional connectivity. The authors computed model-free metrics with mutual information, phase-locking value and phase-transfer entropy across carefully chosen Brodmann-area electrodes, yielding 16 $\times$ 16 adjacency matrices per trial. Treating these matrices as images, a subject-specific CNN classified the low, medium and high load with peak accuracies exceeding 97%. This high performance of connectivity maps emphasise that fatigue does not merely alter local oscillations but reconfigures network-level interactions among frontal, parietal and insular nodes.

Collectively, these studies show a clear progression in EEG-based workload and fatigue research. We see that early reliance on single-domain features has evolved into multi-stream deep models that can exploit spatial, spectral, temporal and network information. Accuracy has climbed from mid-80 % in cross-task SVM pipelines to above 95 % in modern end-to-end networks, with interpretability techniques making sure that the performance gains are physiologically grounded. These works have influenced the present thesis, which utilizes source-localised signals and a CNN-BiLSTM model and SHAP explanations to further close the gap between real-time classification and neuroscientific validity in mental-fatigue monitoring systems.

Chapter 4

Methodology

This section describes the methodology used in this study for mental fatigue assessment using EEG signals. It covers the design of the experimental paradigm in depth, advanced signal preprocessing techniques, including source localization with sLoretta, and a deep learning classification model that combines convolutional neural networks (CNN) and bidirectional long short-term memory (BiLSTM) networks. The proposed framework aims to capture spatial and temporal patterns in EEG data to effectively classify different mental fatigue levels.

In this chapter, a 2-back task paradigm will be explained, that was used to induce different levels of mental fatigue. This paradigm is well known to impose sustained cognitive load and engage working memory. EEG signals were recorded during the performance of the task and at rest to capture neural activity under different conditions of fatigue. These data also underwent preprocessing to enhance signal utility and ensure that meaningful physiological features could be extracted. The source localization was the most crucial step using the sLoretta that mapped the EEG signals obtained on the scalp to underlying cortical regions. Such approaches allow greater spatial resolution but more importantly, higher interpretability as neural activity in connection with specific cognitive functions and fatigues can be discriminated. The EEG signals were then segmented into epochs of 4-second (each with 1024 samples) that are appropriately considered for better analysis as used with the CNN-BiLSTM model. It is very important to note that signals already had artifact removal performed before preprocessing, so it eliminated the need for these extra steps to remove the artifacts.

The hybrid CNN-BiLSTM model that was used will also be explained, a model that leverages the strengths of both architectures the CNNs for spatial feature extraction across EEG channels and BiLSTMs for modeling the temporal dynamics of neural activity. The model automatically learns discriminative patterns from preprocessed EEG data without using handcrafted features. It further enhances the spatial features by incorporating source-localized data to improve the accuracy of fatigue classification. This section further gives an explanation of explainability of SHAP (Shapley Additive Explanations) values that are incorporated to understand the model's predictions. SHAP provides insight into

how different features contribute to the classification of mental fatigue states, thus increasing model interpretability and decision credibility. This approach ensures physiologically relevant neural patterns, learned by the model, are consistent with a solid body of knowledge pertaining to the mechanisms of mental fatigue.

4.1 Working Memory and the n-Back Task

Working memory is the core component of human cognition which empowers the temporary storage and manipulative functions of information needed in executing complex cognitive tasks like reasoning, problem-solving, and decision-making. It is a cognitive workspace, where individuals can hold and update information relevant to ongoing tasks while simultaneously processing new inputs. While distinguished from short-term memory, which mainly holds information for brief periods of time, working memory deals with the active manipulation of that information to govern behavior. Working memory is utilized internally while calculating math problems in one's head to remember intermediate results as one processes subsequent steps of the problem [28].

4.1.1 N-back task

The n-back task is one of the most widely used paradigms for studying working memory in cognitive neuroscience. It involves the sequential presentation of stimuli (e.g., letters, numbers, images) and requires participants to determine whether the current stimulus matches the one presented n steps earlier. The value of n can be adjusted to manipulate task difficulty and cognitive load, with higher n levels placing greater demands on working memory [29]. In a standard n-back task, stimuli are presented one at a time in a continuous stream. Participants must respond (e.g., via button press) when the current stimulus matches the one presented n items earlier. For example in a 1-back task, participants compare the current stimulus to the one immediately preceding it. On the other hand in a 2-back task, participants compare the current stimulus to the one presented two steps earlier. Also in a 4-back task that we can see in Fig 4.1, participants compare the current stimulus to the one presented four steps earlier and so on.

N-back task can also be adapted in several ways to target specific cognitive processes or experimental objectives. One example of this is that the stimuli can be auditory (e.g., spoken words) or visual (e.g., shapes, letters). Multimodal versions combine both modalities to assess integration across sensory systems. There are also the dual-task versions require participants to perform a secondary task simultaneously, further increasing cognitive load. These tasks can use verbal stimuli (e.g., letters, words), numerical stimuli, spatial locations, or abstract shapes, depending on the research focus.

One more import aspect of the n-back task is that it is particularly effective for assessing cognitive load because it allows for precise control over task difficulty by manipulating the value of n. Increasing n requires participants to hold more items in working memory

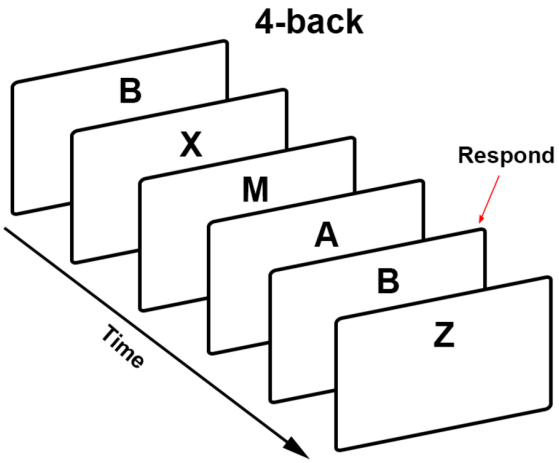


Figure 4.1: Example of 4-back task [4]

while simultaneously comparing incoming stimuli to past ones, thereby taxing attentional resources and executive control. Furthermore, the n-back task's design enables real-time measurement of performance through metrics such as response accuracy, reaction time, and error rates. These behavioral indicators provide a direct link between cognitive load and task performance, which can be further complemented by neural measures such as EEG. N-back task is particularly well-suited for studying mental fatigue due to its ability to systematically manipulate cognitive load over time. Mental fatigue arises from prolonged engagement in cognitively demanding activities, leading to a decline in performance and attentional control. By increasing the difficulty or duration of the n-back task, researchers can induce fatigue and measure its effects on both behavioral and neural responses.

Higher n levels in the n-back task put higher demands on working memory, requiring participants to use more cognitive resources to maintain performance. During fatigue, the ability of the brain to maintain these resources decreases, and clear differences in task performance become evident. Among the most pronounced behavioral effects are that prolonged performance of the n-back task is associated with prolonged reaction times, reduced accuracy, and higher error rates, particularly at high n levels. The n-back task provides a useful framework for the study of the evolution of cognitive performance in time under sustained workload conditions.

4.2 Dataset description

The dataset that we utilized in this thesis was collected to investigate the effects of mental fatigue on working memory with EEG signals. The participants were selected carefully, and the experimental protocol was designed to evoke neural activity under both rested and fatigued conditions, providing a good foundation for analyzing the influence of sleep deprivation on cognitive function. The experimental setup and preprocessing for cleaning the data were done before acquiring the dataset for the work in this thesis.

4.2.1 Participants and experimental design

For the experiment 22 healthy participants (9 females, mean age 27.3 ± 4.1 years) were recruited from the 401 General Military Hospital of Athens, comprising doctors and staff members. All the participants were prescreened to make sure that they had no history of sleep disorders, mental illnesses, ADHD, or long-term medication use, as these factors could affect the results. All participants also reported normal or corrected-to-normal vision. Ethical approval for the study was obtained from the institution's Review Board, in compliance with the Declaration of Helsinki. Also the written informed consent from all participants was signed before the commencement of the experiment, to make sure that they participate voluntarily and they understand the study's procedures.

Regarding the experiment a visual 2-back working memory task was employed to evaluate the cognitive effects of mental fatigue. This 2-back task is a reliable method for imposing sustained working memory demands. Each participant completed the task twice, once before their on-call shift (the Rested condition) and once after their on-call shift (the Fatigue condition), which lasted up to 28 hours with minimal or no sleep. This protocol allowed for a direct comparison of neural activity under rested and fatigued states, isolating the effects of mental fatigue induced by sleep deprivation. During this 2-back task, participants were required to compare the current visual stimulus with one that was presented two trials earlier. The stimuli were images displayed in one of the four corners of the screen, and participants had to evaluate the stimulus based on both content and location. Using a response box, participants indicated one of four conditions:

1. Same image and same location
2. Same image but different location
3. Same location but different image
4. No similarity (different image, different location)

The specific task consisted of 72 trials, with the four conditions being balanced across the session. Each trial lasted approximately 5 minutes, with the visual stimuli displayed for 3.5 seconds and with a 1-second fixation cross. To ensure that participants fully understood the task, they completed practice trials before the actual EEG recordings. This preliminary step minimized the likelihood of errors due to misunderstanding the task instructions.

4.2.2 Data acquisition and preprocessing

EEG data were collected by means of the Biosemi Activetwo System comprising of a total of 64 channels according to the standards of the 10-20 international electrode positioning system [5]. The system was beneficial since it provided sufficiently high spatial resolution to enable recording from specific brain areas linked to working memory as well as to mental fatigue. 512 Hz was the EEG sampling rate that was adopted for this

study as it was enough to ensure that the temporal variations of neural activities were preserved. Also, the bilateral EOG signals were captured using bipolar electrodes placed at the sides of the eyes in order to track and remove any eye movement induced noise. Specific measures were taken with regard to the set up in order to limit the impact of external noise and movement on the EEG data recorded. The measures ensured that the EEG signals were reliable and consistent across different subjects. Raw EEG Signals were deposited into extensive preprocessing to wipe them out and to promote their quality as well their suitability of use for subsequent analytical purposes.

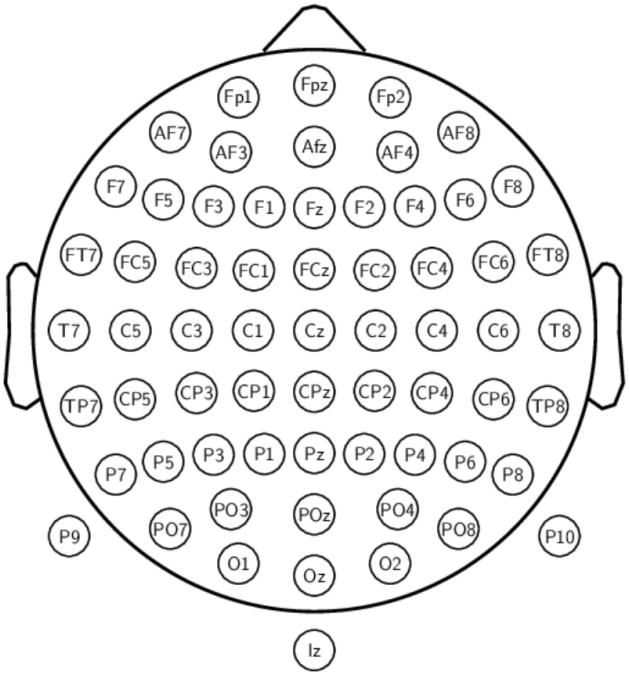


Figure 4.2: Electrode placement of the experiment [5]

These preprocessing steps, that were performed prior to the acquisition of data for this study, included:

1. Down sampling and band pass filtering. This involves reducing the raw eeg database from 512 hz to 256 hz to minimize computational difficulties while still maintaining the essential temporal emblems. An appropriate cut-off frequency between 1Hz and 40 Hz was implemented since this range was sensitive to cognitive and fatigue processes. In addition to lowering frequency drifts, these filtering strategies did away with high frequency noises.

2. Artifact Correction Using ICA: Eye blinks, muscle movements, or any other non-brain activities were separated and removed from the EEG signals using ICA. This method finds application in Reconstruction of EEG signals through Independent Component Analysis (ICA), where the components of these EEG signals are transformed into independent variables which can then be used to ascertain and eliminate

together with EOG and EMG signals [30]. The independent components are primarily easy to classify into EOG and EMG noise through their topographic map and signal shape. However, there are also cases where it is difficult to distinguish what type of signal is the independent component where the independent component has an EOG signal topographic map shape, whereas it is most likely an EMG signal. So the classified EOG and EMG signals with a big probability that they are the respective signals are rejected.

3. Segmentation: The procedures involved divided the EEG signals into 4-second periods a well performing segmentation size [31] or referred to as segments of 1024 sampling points. EEG signals are non-stationary and change rapidly over time. By segmenting the data, these dynamic fluctuations are captured more effectively. Also, the temporal information of the EEG is better utilized. Each segment was shifted along the time axis with respect to the preceding stimulus along the time axis with respect to the preceding 100 ms stimulus, in order to perfectly minimize inter-trial differences. The first 0.5 seconds after every stimulus presentation was disregarded in the analysis in order to decrease the possibility of capturing effects associated with a transient vision stimulus.

4. Electrode Selection: In order to ensure symmetry with regards to the analysis, Iz electrode was eliminated which left 63 electrodes to be analyzed afterward. This arrangement was favorable as it made sure that no brain regions became deficient or over represented in the activity under study.

5. Trial Selection and Labeling: The only trials which were kept were the ones that resulted from correctly answering the task thus removing any cognitive activity that was not related to the use of working memory. This provided a total of 2047 trials out of which 1050 originated from the Fatigued condition while 997 came from the Rested condition.

To conclude every trial of the dataset is saved as a 63×1024 matrix, where the rows correspond to EEG signals from 63 channels, and the columns represent time points sampled at 256 Hz. This format ensures that each trial contains good spatial and temporal information for the model to make the necessary associations between the mental states that the participants had in the experiment and have a good recognition accuracy. The dataset is also balanced across fatigue states, ensuring that the classification model can learn and distinguish patterns related to rested and fatigued conditions more effectively.

4.3 Source localization

Source localization is a rather complex and advanced technique which is applied in EEG for the purpose of estimation the cortical areas from which the recorded electrical activities are derived [32]. This resolution is low but the temporal resolution is highly

appreciated. There is a restriction in the context of recording electrical signals because of the loss of the characteristics of such signals that are moving through brain, skull, scalp and the EEG scalp electrodes. The first approach to solving this loss of information concerns the mathematical reconstruction of the neural generators focusing on the sources of EEG signals. This helps pinpoint the cortical regions that are responsible for specific cognitive or neural functions. As stated above, the EEG signals are produced by the integration of the postsynaptic potentials of large number of neurons. These signals originate in the cortex and consider propagating through different conductive media such as, CSF, skull, scalp and electrodes. These electrodes measure the signals but they do not analyze the waves exactly. As a result, it is observed that a single EEG electrode is in fact a large spatial integrate of many cortical regions and thus the exact origin of these signals cannot be directly confirmed from the scalp recordings.

In the effort to localize the existing current sources inside the brain based on the potentials and their distribution recorded by EEG electrodes on the scalp, the potential combinations of the current sources that generate such potentials and distributions and that are captured by the pre-defined number of electrodes might be infinite [33]. This case is known as the inverse problem, which could be resolved by introducing constraints based on the anatomical and physiological rules that control the generation and propagation of the current sources. Various proposals have been introduced to attempt to solve the matter with improvements and justifications provided for the proposed models and the introduced technical approaches. Hence, this review defines the source analysis models used in dipole source localization and distributed source localization that prevail currently in the domain.

Distributed source localization estimates the 3-dimensional structure of the brain into many lattice points, typically over 5,000, and offers a model where current dipoles placed in each lattice are distributed in their respective strengths. Hence, it is not required to determine or assume. The number of existing sources is high. However, the number of lattice points of the distributed source model far outweighs the number of electrodes measured on the scalp, therefore, one has to deal with an inverse problem. In the head model, the location of the lattice points is restricted to gray matter and hippocampus, using either individual or template MRI scans, and the constraints provided by this anatomical information reduce the number of variables. Additionally, under the distributed source model, there is an minimum norm (MN) methodology which states that a distribution is optimal when it has a minimum total energy of all the current sources [34]. For this purpose, the MN solution tends to stay near the scalp electrodes and thus may fail to recognize current sources with larger depths in the brain. To counteract this disadvantage, a depth-weighted MN has been suggested. Notwithstanding, the depth-weighted MN solution exhibits a diminished level of resolution.

At present, distributed source models have garnered significant attention and are evolving in sophistication [35]. The LORETA software employs the Laplacian-weighted MN method, which posits that neuronal activities in close proximity are interconnected where, the distribution of current density is subjected to smoothing. Nonetheless, the LORETA

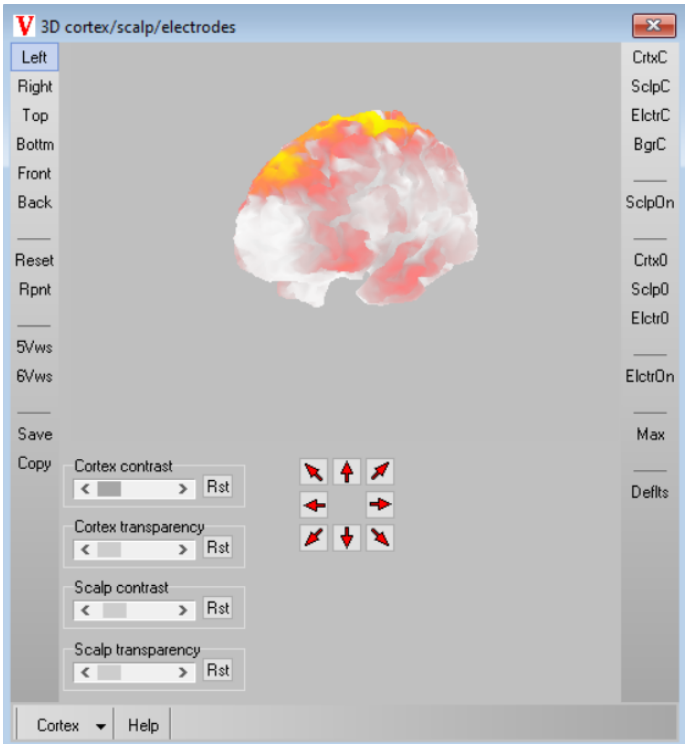


Figure 4.3: 3D cortical view of neural activity across the cortex at specific time point

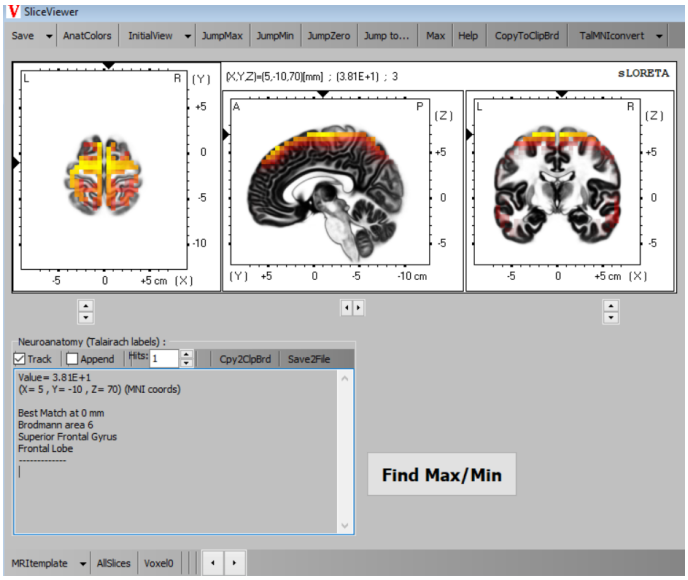


Figure 4.4: Slices of brain activity at specific time point

software is limited by its propensity to generate indistinct and excessively smoothed images. LORETA has been updated to sLORETA (and more recently eLORETA) to compensate for these shortcomings.

We use source localization in our preprocessing pipeline to map scalp-recorded EEG data onto cortical sources via sLORETA. This step enhances spatial resolution by pin-

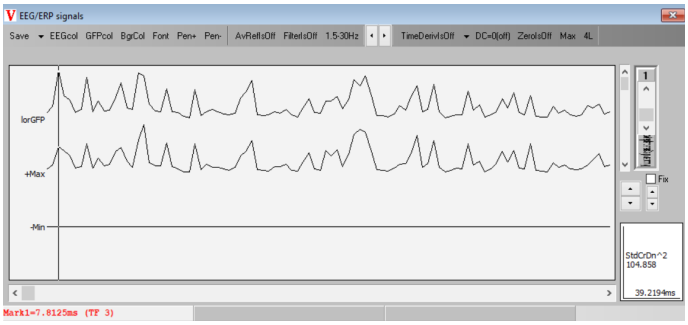


Figure 4.5: EEG signal linked to the localized activity at specific time point

pointing the neural origins of the detected electrical signal. When the source is localised, the whole cortical surface gets partitioned into 80 ROIs, so we are piecing the brain into different sources. This parcellation bundles local activity into relevant anatomical or functional units, compressing the data while retaining vital spatial data. All of the 80 sources are located in a part of the brain with a known neural pattern, giving a refined source to be processed. Then these concatenated sources become the foundation for feature extraction, making sure the EEG signals identifying to the most relevant neural activity for mental fatigue classification are recorded, and the computation is less complicated.

4.4 Standardization

After loading, selecting, segmenting and doing source localization on the data, the next vital step in the preprocessing pipeline is data standardization, and more specifically standard scaling. This process is performed to ensure that EEG data across channels have a common scale, a prerequisite for many machine learning algorithms [36]. Standard scaling is performed by subtracting the mean and dividing by the standard deviation of each feature. Mathematically, this can be represented as follows:

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

- z: The variable z represents the standardized value (or z-score) of the data point x. It indicates how many standard deviations a given data point deviates from the mean of the distribution. A positive z implies that the data point lies above the mean, whereas a negative z indicates that it lies below the mean. This transformation standardizes all data points, allowing them to be compared on a common scale.

- x: The variable x is the raw, unprocessed value of the feature being analyzed. This could represent any measurement within the dataset, such as the height of an individual in centimeters or the temperature in degrees Celsius. The purpose of standardization is to transform x into a comparable metric across features with different scales.

- μ: The mean μ is the average value of the feature across all data points in the dataset. Subtracting the mean μμ centers the data, aligning the feature's distribution around zero.
- σ: The standard deviation σ quantifies the spread or variability of the feature values around the mean. Standard deviation ensures that the data is scaled in a way that reflects its variability. Dividing by σ normalizes the variance of the data to 1, ensuring a uniform scale for all features.

By applying this transformation, the resulting distribution of data points has a mean of 0 and a standard deviation of 1. This technique is particularly useful when dealing with data that have varying scales and ranges. It is important to note that normalization in this sense differs from min-max scaling, which scales the data to a fixed range, usually between 0 and 1. Standard scaling is particularly beneficial in EEG data analysis, as it addresses two important challenges: high inter-subject variability and inherent inter-channel variability in EEG data. With data scaling, the algorithm can learn better from the dataset, as it ensures that no particular feature (or in this case, channel) dominates others due to differences in their scale. Furthermore, it allows us to effectively compare data between different individuals and trials, paving the way for more accurate models of mental fatigue assessment. Thus, standardized scaling of EEG data is an integral pre-processing step that ensures that all data conform to a common scale, promoting fairness and efficiency in subsequent data analysis tasks, particularly those involving machine learning models.

4.5 CNN-BiLSTM hybrid model

This proposed model utilizes the hybrid 2d convolutional neural network (2DCNN) architecture and the bidirectional long-short term memory (BiLSTM) networks to assess with details the mental fatigue of the subjects. This hybrid CNN-BiLSTM model can successfully capture the spatio-temporal information of the EEG signals and utilize the dynamic nature of the signal both from the electrode placements and the signal fluctuations over time. Below the components of the proposed model are explained in detail and its role and design and operational strengths are highlighted in detail.

4.5.1 2DCNN architecture

The 2DCNN component is specifically designed to extract spatial features from the EEG input, represented as a matrix of size 80×1024 (80 ROIs $\times$ 1024 time points). Each channel corresponds to an electrode placed on the scalp, and the time points represent sequential EEG activity. By applying 2d convolutional layers, the CNN identifies patterns in neural activity distributed across the electrodes, such as localized activations or inter-channel relationships that are indicative of mental fatigue.

Each 2D convolutional neural network (2DCNN) layer is designed to extract spatial features from input data, such as EEG signals represented as a 2D matrix. It works by

applying convolutional filters (kernels) over the input matrix to detect localized patterns, such as correlations between adjacent electrodes or specific spatial activity. The convolution operation involves sliding a kernel of predefined size ($k \times k$) across the input matrix and computing the dot product between the kernel weights and the overlapping input region. Mathematically, the operation can be expressed as:

$$Y(i,j) = \sum_{m=1}^{k} \sum_{n=1}^{k} X(i+m, j+n) \cdot W(m,n) + b$$

where:

- $Y(i,j)$ is the output feature at position (i,j),
- $X(i+m, j+n)$ represents the input matrix values covered by the kernel at position $(i+m, j+n)$,
- $W(m,n)$ is the kernel weight matrix,
- b is the bias term.

The resulting feature map Y is then passed through a tanh (hyperbolic tangent) activation function, defined as:

$$f(x) = \tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$$

This introduces non-linearity, enabling the network to model complex spatial patterns. The feature maps produced by the 2DCNN layer highlight important spatial relationships, such as localized neural activity or inter-channel dependencies in EEG data. One key feature of the CNN architecture is the use of progressively larger kernel sizes in the convolutional layers. The model employs kernels of sizes 3×3, 5×5, and 7×7 in successive layers to capture features at varying spatial scales. This progressive increase in kernel size allows the CNN to extract both fine-grained and global spatial features:

- 3×3 kernel (first layer): The smallest kernels are used in the initial convolutional layer to focus on localized features in the data. These features might include sharp changes in activity between neighboring electrodes or small-scale patterns that are critical for identifying the early signatures of fatigue-related neural activity.
- 5×5 kernel (second layer): The second convolutional layer expands the receptive field to capture regional patterns, aggregating information from clusters of electrodes. For example, this layer might identify activity patterns within specific cortical regions, such as the frontal or parietal areas, which are often implicated in working memory and cognitive load.
- 7×7 kernels (third layer): The largest kernels are employed in the final convolutional layer to detect global spatial features. These features represent widespread neural activity patterns or inter-regional correlations that become prominent during fatigue.

Also each 2DCNN layer is followed by a MaxPooling layer, which is a sampling reduction function that is usually applied after convolution operations in a CNN. The primary purpose of the MaxPooling layer is to progressively reduce the width, height and, in the case of the 2DCNN, depth of the input representation (in our case reducing temporal and spatial dimensions). This helps to reduce the amount of parameters and computations in the network, thus controlling overfitting. A MaxPooling layer works by sliding a 2D window in the input volume and, for each window, extracts the maximum value. In essence, it selects the most prominent features within the window, discarding the least prominent ones. This feature provides the network with a certain level of invariance to small changes or distortions in the input data, allowing it to focus on the most prominent, high-level features.

4.5.2 BiLSTM components

The BiLSTM element attempts to identify the temporal dependencies of EEG signals [37]. While CNN is interested in the spatial associations and temporal features, the BiLSTM layers look at how those spatial characteristics change with time, which helps in detecting the onset of mental exhaustion. In contrast to the more conventional RNNs, BiLSTMs encode input sequences both forward and backwards, so that the model can work on past and future simultaneously. Each LSTM unit of the BiLSTM network contains a set of gates the input gate, forget gate, and output gate controlling the flow of information in the network. These gates allow BiLSTM to store or remove data only when it's needed, a feature especially helpful when dealing with long term dependencies on sequential data. This gating process is essential for the interpretation of EEG data, because fatigue-related patterns can arise over long periods of time. Because the BiLSTM is bidirectional, this makes it useful for modelling temporal relationships using the following technigues:

- Forward Pass: The forward pass processes the input sequence from the beginning to the end, capturing dependencies where earlier time points influence later activity. This is useful for identifying fatigue-related trends that develop over time.

- Backward Pass: The backward pass processes the sequence from the end to the beginning, capturing dependencies where future time points provide context for earlier activity. This bidirectional processing is particularly advantageous for EEG signals, where both past and future neural dynamics contribute to the observed patterns.

The BiLSTM component in our model consists of two stacked layers designed to process sequential data and capture temporal dependencies in the spatial features extracted by the CNN layers. The first BiLSTM layer takes the reshaped output from the final 2DCNN layer as its input. This input, that is formatted as a sequence, allows the BiLSTM to analyze how the extracted spatial features evolve over time. The first layer processes the sequence in the two directions the forward pass, which examines the data from the beginning to the end, and the backward pass, which processes the data in reverse. This

bidirectional processing enables the network to consider both past and future contexts, ensuring a comprehensive understanding of the temporal dependencies in the data. The hidden state outputs generated by this layer encapsulate the temporal relationships across the input sequence, laying the foundation for deeper temporal modeling.

The second BiLSTM layer extends the representations generated by the initial layer. With the help of these intermediate temporal features, the second layer polishes and enhances the learned temporal characteristics. This stacked configuration enables the model to identify more elaborate and structured temporal characteristics which can be useful in differentiating between various fatigue stages. This layer's output is a sequence of concatenation of the forward and backward hidden states for each time step, which gives a lot of information about temporal dynamics.

Every BiLSTM layer includes 128 hidden units, which specifies the size of the hidden state variables. This size is enough for the network to have enough capacity to capture the essential temporal features while at the same time being computationally efficient. Also, a dropout layer is used before the two BiLSTM layers to avoid overtraining. In dropout, a portion of neurons are switched off randomly during the training process in order to make the model more robust to data that it has not seen before. The output of the BiLSTM part is the representation of the temporal features obtained from the EEG data. This output is then fed into fully connected layers in order to categorise the different mental fatigue states using the temporal information as well as the spatial patterns identified by the CNN. This architecture of the BiLSTM component with the stacked layers and the bidirectional connection is capable of capturing the temporal evolution of neural activity, thus enabling the model to learn the complex patterns associated with mental fatigue.

4.5.3 Fully connected layers

The extracted features are already a flattened vector after the pass on the second BiLSTM layer so it is then passed through a fully connected layer (Dense layer) with an activation function 'relu' and 1024 output neurons and its output passes through a a fully connected layer with 128 output neurons. The final Dense layer works as the output layer, using a 'softmax' activation function to generate the probabilities for each of the two classes of mental state. In the following Table 4.1 we can see in detail the layers of the proposed model.

Layer (type)	**Output Shape**	**Kernel Size**	**Activation Function**
InputLayer	$80 \times 1024 \times 1$	-	-
Conv2D	$80 \times 1024 \times 32$	$k \times k$	Tanh
MaxPooling2D	$40 \times 512 \times 32$	-	-
Conv2D	$40 \times 512 \times 64$	$k \times k$	Tanh
MaxPooling2D	$20 \times 256 \times 64$	-	-
Conv2D	$20 \times 256 \times 128$	$k \times k$	Tanh
MaxPooling2D	$10 \times 128 \times 128$	-	-
Dropout	$10 \times 128 \times 128$	-	-
TimeDistributed_Flatten	10×16384	-	-
TimeDistributed_Dense	10×128	-	ReLU
BiLSTM	10×256	-	-
BiLSTM	256	-	-
Dense	1024	-	ReLU
Dropout	1024	-	-
Dense	128	-	ReLU
Dense	2	-	Softmax

Table 4.1: Summary of the Model Layers

4.6 Model interpetability with SHAP values

A crucial aspect of this study is the integration of SHAP (Shapley Additive Explanations) values to enhance the interpretability of the CNN-BiLSTM model used for mental fatigue detection. While deep learning models excel at extracting patterns and achieving high classification accuracy, their complexity often makes them "black boxes", limiting insight into how predictions are made. SHAP values address this issue by providing a transparent, mathematically rigorous framework to explain the contribution of each input feature to the model's output. By applying SHAP to the CNN-BiLSTM model, this study bridges the gap between performance and interpretability, offering both accurate predictions and physiologically meaningful insights.

SHAP values are grounded in cooperative game theory and extend the concept of Shapley values, which were originally developed to fairly distribute rewards among players in a game. In the context of machine learning, SHAP values quantify the contribution of each input feature (e.g., EEG channel and time point) to the model's prediction. Given a model $f(x)$ and an input feature set $x = \{x_1, x_2, \ldots, x_n\}$, the SHAP value for a feature x_i is computed as:

$$\phi_i = \sum_{S \subseteq \{x_1,\ldots,x_n\} \setminus \{x_i\}} \frac{|S|!(n - |S| - 1)!}{n!} [f(S \cup \{x_i\}) - f(S)]$$

Where:

- ϕ_i: The SHAP value for feature x_i, representing its contribution to the model's prediction.
- S: A subset of all features excluding x_i.
- $f(S)$: The model's prediction when only the features in subset S are included.

In our framework, each EEG segment after source localization is represented as a series of activities in 80 cortical ROIs. The CNN-BiLSTM model then uses these ROI-specific time signals to classify whether the participant is "rested" or "fatigued." While this pipeline yields high accuracy, SHAP serves as a post hoc explainer to reveal which ROIs and temporal windows most strongly help with the model's classification. Firstly we treat each ROI as a feature, capturing its spatiotemporal activity. In practice, the CNN part of the network captures spatial relationships across ROIs, while the BiLSTM uncovers temporal dependencies. After the model is trained, we extract predictions for each test instance. Then we employ a model-agnostic SHAP method specifically the GradientSHAP that uses gradient-based approximations. Since computing exact Shapley values can be extremely computationally expensive, these approximations strike a practical balance between speed and fidelity of explanations. After that all of the test instances local explanations are generated, SHAP yields a vector of 80 attributions, one for each ROI along with the model's baseline logit or probability. ROIs with positive SHAP values

push the model's output closer to the "fatigued" side, whereas negative values push it toward "rested." The magnitude of the value indicates how influential that ROI is relative to others. Finally all local explanations can be aggregated across multiple trials and subjects to uncover consistent patterns. For instance, if frontal ROIs generally exhibit higher positive SHAP values in fatigued epochs, it supports neuroscientific findings about frontal-lobe engagement under sleep deprivation or high cognitive load.

Chapter 5

Experiments and Results

5.1 K-fold cross-validation

In this chapter the experimental procedure and the results obtained are explained and compared with the results of similar works. The experimental procedure incorporates a popular, widely accepted methodology in machine learning and statistics the k-fold cross-validation. However, before applying this procedure, the dataset is first divided into two distinct subsets, a training set and a test set. This initial separation is an integral part of the machine learning process, ensuring that the model, while learning from the training set, is validated on unseen data (test set) that played no role in the training phase. The training phase uses the concept of k-fold cross-calidation, a powerful resampling technique used for model evaluation and selection. The following is how this technique works:

1. The training dataset is divided into ¨K¨ equal folds or subsets. Each of these subsets has an equal chance of being used as a validation set while the model is trained on the remaining subsets.

2. This procedure is performed in a loop for ¨K¨ iterations. In each iteration, one of the ¨K' subsets is used as the validation set and the remaining ¨K-1' subsets form the training set.

3. The model is trained on the 'K-1' training subsets and then validated on the validation subset that was reserved. The performance is evaluated using the accuracy metric, which is then stored.

4. Steps 2 and 3 are repeated ¨K' times until each unique subset is used once as a validation set. This ensures that every observation from the original training dataset has a chance to be validated.

5. Calculate the average validation error across all the 'K' trials. This average error serves as the overall performance metric of the model.

The underlying strength of k-fold cross-validation lies in its integrated nature, significantly reducing the bias and variance associated with a single experimental run. It ensures

that each data point is found once in a validation set and ¨K-1' times in a training set, providing a more accurate and reliable measure of model performance, especially when the size of the dataset is limited. In this thesis, 10-Fold Cross-Validation is used, which is the same as the amount of subjects where the EEG data were collected. Also, after each model is trained and validated with the corresponding fold, the final evaluation is performed with the test set, which has not participated in the training at all, and stored to calculate the final average accuracy value from all 10 models.

5.2 Optimizer and loss function

The model is drawn with the Adam optimizer with a learning rate of 0.0001 and a decomposition of 0.00001, using categorical cross-entropy as the loss function and precision as the metric. The Adam optimizer (Adaptive Moment Estimation) is a stochastic gradient-based optimization algorithm. It is a popular choice due to its efficiency and low computational resource requirement. Adam calculates adaptive learning rates for different parameters, which makes it particularly effective when dealing with sparse or noisy data. It achieves this by estimating the first and second moments of the gradients to adapt the learning rate, hence it is called adaptive moment estimation.

The learning rate determines how much the model will change in response to the estimated error each time the model weights are updated. The choice of learning rate for the Adam optimizer can be important, as it controls the step size at each iteration while moving towards the minimum of the loss function. A lower learning rate can lead to more accurate convergence towards the minimum of the loss function, at the cost of convergence speed. In this model, Adam is used with a learning rate of 0.0001 and an attenuation rate of 0.00001. .The learning rate determines the size of the step at which the optimizer moves towards the minimum of the loss function, and the decay rate slowly decreases the learning rate over epochs, allowing the model to learn more efficiently.

Categorical Cross-Entropy Loss is a loss function often used in machine learning models for multi-category classification problems. The loss indicates how far a model's prediction is from the actual data. This loss is commonly used in models where the output is a probability distribution. It is also used when we have one-hot coding. In one-hot coding, each categorical value is converted to a new categorical column and assigned a binary value of 1 or 0. Each integer value is represented as a binary vector. All values are zero and the index is marked 1.

The mathematical formula for the categorical cross-entropy loss function for N classes is as follows:

$$L = -\frac{1}{N}\sum_{i}\left[y_i \log(y_i')\right]$$

Below are the components of the formula:

- N: The total number of classes.

- y_i: This is the true label for class i, often represented as a binary value, where 1 indicates the correct class, and 0 for all others (one-hot encoding).
- y_i': This is the predicted probability that an example belongs to class i, as provided by the model's output (usually a softmax function).
- log: This is the natural logarithm function.
- $\sum$: This represents the summation over all classes i.

The term $-1/N$ is the average across all N classes. In essence, categorical cross-entropy loss calculates the difference between two probability distributions the true distribution (the coded labels using the one-hot method) and the predicted distribution (the output of the softmax function in the model). The logarithmic function provides a large penalty for predictions that are confident but incorrect and a small penalty for predictions that are correct. The negative sign up front ensures that the loss is positive, as logarithm values for numbers between 0 and 1 are negative.

5.3 Metrics

Evaluation metrics play an important role in machine learning as they provide a quantitative measure of the model's performance, thus providing a clear understanding of its strengths and weaknesses. This section mainly focuses on accuracy and confusion matrix, two key metrics in the context of fatigue assessment based on EEG signals.

5.4 Accuracy

Accuracy is a simple and intuitive metric, representing the fraction of predictions that our model gets right. Mathematically, accuracy is defined as the ratio of correctly predicted instances to the total instances in the dataset. The mathematical equation for accuracy is:

$$\text{Accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{\text{True Positives} + \text{False Positives} + \text{False Negatives} + \text{True Negatives}}$$

Where:

- True Positives (TP):These are cases where we predicted "yes" (positive), and the actual value was also "yes."
- True Negatives (TN): These are cases where we predicted "no" (negative), and the actual value was "no."
- False Positives (FP):These are cases where we predicted "yes," but the actual value was "no."

- False Negatives (FN):These are cases where we predicted "no," but the actual value was "yes."

In this equation, the numerator represents the correct predictions (positive and negative) and the denominator is the sum of all the outcomes, correct or incorrect. Thus, precision gives us a ratio of the correctly predicted observations to the total number of observations. In this paper, however, we use the average accuracy, which is the sum of all accuracies from all models (folds) divided by their n number: Average Accuracy $= \frac{\text{Acc}_1+\text{Acc}_2+\cdots+\text{Acc}_n}{n}$. Thus with the average accuracy we can effectively evaluate the overall performance of all models.

5.5 Confusion matrix

The confusion matrix, also known as the error matrix, is a specific matrix layout widely used in machine learning and statistics to visualise the performance of an algorithm. It is particularly useful for supervised learning problems. Each row in the table represents instances in a predicted class, while each column represents instances in an actual class. The name comes from the fact that it makes it easy to see if the system is 'confusing' two classes (it usually misidentifies one as other).

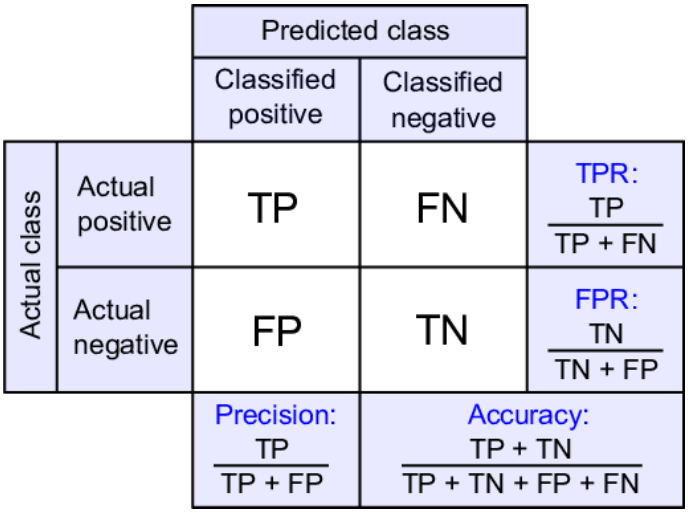


Figure 5.1: Schematics of the confusion matrix for the binary classification problem including definitions of basic terms used in the assessment of model's performance [6].

The following is a brief overview of the basic components of a confusion table: - True Positives (TP): these are the correctly predicted positive values for each class, meaning that the class was accurately predicted as the true outcome. These values are located along the main diagonal of the table. - True Negatives (TN): For any given class, these are the correctly predicted negative values, meaning that the other instances of the class were correctly identified. - False Positives (FP): also called Type I errors, they occur when a class is incorrectly identified as the target class. - True Negatives (FN): Also known as

Type II errors, they represent cases where the target class was incorrectly classified as another class.

The confusion matrix allows a variety of metrics such as precision, accuracy, recall (also known as sensitivity), and F1 score to be calculated for each class, and these metrics can be combined or weighted in various ways to provide a single overall measure of model performance. It is particularly useful in cases where the classes are unbalanced, i.e. where some classes have many more instances than others, which is a common situation in real-world datasets.

5.6 Results

In order to evaluate the performance of the CNN-BiLSTM model, a rigorous 10-fold cross-validation method was adopted in this thesis to generate the classification results. The model yielded an average accuracy of 91.55% with a standard deviation of 2.67% for each fold of the cross-validation. This indicates that the model performs fairly consistently and with a low level of variation, and thus it is considered to be quite accurate in identifying the different mental fatigue states from EEG data. The database was split into ten equal parts, and the K-fold cross-validation was employed, where each set was used for validation while the remaining nine were for training. This approach makes sure that the model is tested on every part of the dataset making it less likely to overfit and thus provide a better measure of performance. The high average accuracy indicates the ability of the model in producing correct results even when applied to different parts of the data set.

The standard deviation of 2.67% shows the consistency of the model's performance throughout the folds, and the accuracy did not fluctuate much. This shows that the CNN-BiLSTM architecture is capable of identifying the key features of the EEG signals which represent the transition from rested to fatigued mental state. Consequently, it is seen that the CNN-BiLSTM model has a good classification performance for the 10-fold cross-validation with an accuracy of 91.55%. This result supports the model's potential to effectively model EEG data and identify mental fatigue states, which forms a strong base for its utilization in neuroscience study as well as real-life applications.

The normalized confusion matrix with percentages shown in Fig 5.3 helps to easily comprehend the performance of the model on different classes. It is observed that the model was able to identify 93% of the Fatigue samples as Fatigue and 90% of the Rest samples as Rest. The high percentages along the diagonal show that the model performs well in differentiating the two classes. Nevertheless, it also shows low misclassification rates; 7% of the Fatigue samples were classified as Rest and 10% of the Rest samples were classified as Fatigue. This suggests that in some cases the neural signatures of fatigue and rest may be confused especially in the transitional or mild fatigue stages. Remarkably, the model performs better in predicting Fatigue (93%) than Rest (90%), which means that features of fatigue in EEG data may be easier to identify or more constant.

The confusion matrix without normalization shown in Fig 5.2, however, gives the

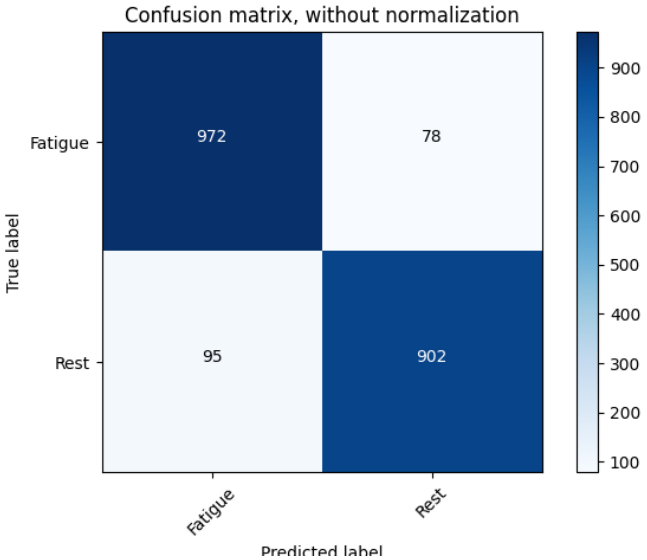


Figure 5.2: Confusion matrix for 10-fold cross-validation experiment

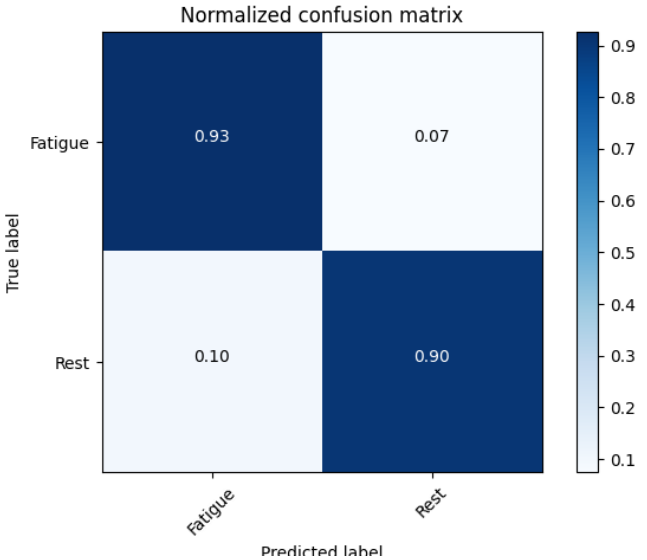


Figure 5.3: Normalized confusion matrix for 10-fold cross-validation experiment

exact number of the correct and wrong classifications making it easier to understand the specific characteristics of the classification problem. Out of 1020 Fatigue samples it correctly classified 972 as Fatigue and 48 as Rest while out of 1020 Rest samples it correctly classified 902 as Rest and 118 as Fatigue. This indicates that the model performed well across both categories with a relatively low error rate. Furthermore, the equally divided data set between the Fatigue class and the Rest class makes it impossible for the model to tend towards one class more than the other.

These two matrices in comparison demonstrate how efficient and coherent the CNN-BiLSTM model is. The normalized matrix helps to understand the proportional accuracy

of the model and makes the class-wise performance more clear while the unnormalized matrix gives an idea about the actual scale of correct and wrong predictions. Taken together, they show that the model is very precise in identifying the states of mental fatigue with minimum chances of misclassifying the states. However, the slightly higher error rate for the Rest class shows the potential for the model to be improved, especially where there are mixed states of being rested and having mild fatigue, for instance. These matrices establish the strong performance of the suggested CNN-BiLSTM approach for managing EEG in the course of mental fatigue. Such performance with high level of accuracy and without big differences between the classes make this structure to be a potentially useful tool and may have its applications in further researches in cognitive neuroscience and general monitoring of brain fatigue.

5.7 Comparative analysis

In comparing our CNN-BiLSTM approach to similar EEG-based methods for mental fatigue detection, several key aspects come to light. First, the protocol,like many others relies on cognitively demanding tasks (here, an N-back paradigm) to induce fatigue, though the exact task variants differ among studies. For example, Xing et al. [24] used both an N-back and a mental arithmetic task, emphasizing a cross-task scenario in which models are trained on one task and validated on another, obtaining around 84.5% accuracy. Their method underlines the inherent challenges of generalizing across different tasks, particularly when using classical fuzzy-entropy features and SVM classifiers.

On the other hand, approaches such as Xu et al. [23] and Su et al. [26] illustrate two distinct success routes, one that uses well-crafted features plus a robust ensemble classifier (Xu et al. using relative band power, fuzzy entropy, and XGBoost, achieving 92.39% accuracy), and another that leverages deep neural architectures (Su et al.'s CNN-LSTM, yielding 97.12% on a three-level fatigue problem). While these methods differ significantly in preprocessing (wavelet denoising vs. simpler filters) and channel counts, they confirm that carefully engineered pipelines can detect nuanced fatigue-related patterns. Karim et al. [38] similarly adopt a smaller CNN (EEGNet) for binary fatigue detection, reaching 88.17%. Their emphasis on portability and real-time practicality with a lightweight 4-channel setup that indicates an emerging trend of wearable EEG solutions, albeit at a modest cost in classification accuracy.

Our own CNN-BiLSTM pipeline joins this deep-learning trajectory: it automatically learns and refines temporal-spatial EEG features, culminating in a strong average accuracy of 91.55% under 10-fold cross-validation for binary (fatigue vs. rest) classification. Comparing to Zhang et al. [25], whose two-stream network (TSN) learns from both spectral and temporal topographic maps and achieves 91.9% for three classes, our technique is similar in fusing different EEG characteristics (temporal and convolutional). Yet, ours maintains a slightly simpler architecture (CNN + BiLSTM) that still robustly captures fatigue-relevant patterns, as evidenced by our confusion matrices (93% accuracy in detect-

ing fatigue, 90% for rest). Minor performance disparities often stem from differences in task designs, sample sizes, EEG channels, artifact removal, and whether the study uses two or three fatigue levels. Still, all of these works confirm that advanced architectures, especially ones blending CNNs with recurrent modules or parallel streams that consistently surpass the earlier hand-crafted and traditional classifier pipelines.

Reference	**Methodology**	**Classes**	**Accuracy**
Xing et al. [24]	Fuzzy-entropy + SVM	2-class	84.5%
Karim et al. [38]	EEGNet	2-class	88.17%
Xu et al. [23]	Relative band power + Fuzzy entropy + XGBoost	3-class	92.39%
Zhang et al. [25]	Two-stream network (temporal & spectral) TSN	3-class	91.9%
Su et al. [26]	Wavelet denoising + CNN-LSTM	3-class	97.12%
This thesis	Source localization + CNN-BiLSTM	2-class	91.55%

Table 5.1: Comparison of methods and accuracies for EEG classification tasks.

Overall, our results illustrate that a hybrid deep network, combining convolutional feature extraction with bidirectional LSTM, can achieve robust binary fatigue detection. Not only does it align with contemporary deep-learning methods in accuracy, but it also affirms that synergy between advanced architectures and balanced experimental protocols can reliably discern fatigued vs. rested mental states.

5.8 SHAP values explainability

In order to uncover which brain sources and timepoints predominantly drive the CNN-BiLSTM model's predictions, we employed a two-step procedure that combines source localization with SHAP (SHapley Additive exPlanations). First, each EEG epoch was projected into localized brain regions (ROIs), for example, "Source_30," "Source_33", and "Source_65,", reconstructing cortical signals of interest. Next, these localized signals were segmented by specific timepoints ("Sample_point_159," "Sample_point_775," etc.) to yield a detailed spatiotemporal representation of the EEG data. Each fold of the cross-validation then yielded SHAP values for both training and test samples, thereby indicating how each source/timepoint combination pushed the model's decision toward fatigued or rested. Once all 10 folds were complete, we aggregated the SHAP values for every feature (every "Source_xx_Sample_point_yyy") and computed its mean absolute SHAP score. Ranking these scores revealed which localized timepoints had the strongest overall impact on the model's output. Finally, we selected the top 20 features with the highest mean absolute

SHAP values across folds and visualized them in the beeswarm plot.

In the plot, each row corresponds to a single feature (for instance, "Source_33_Sample_point_777"), while each dot along that row represents the SHAP value for a single EEG instance. The color gradient have a range from blue (low feature value) to red (high feature value), as it indicated on the right. Points that are located on the positive side (to the right of zero) typically show a feature that increases the probability of a fatigued output, and points that are on the negative side (left) suggest that the feature decreases it.

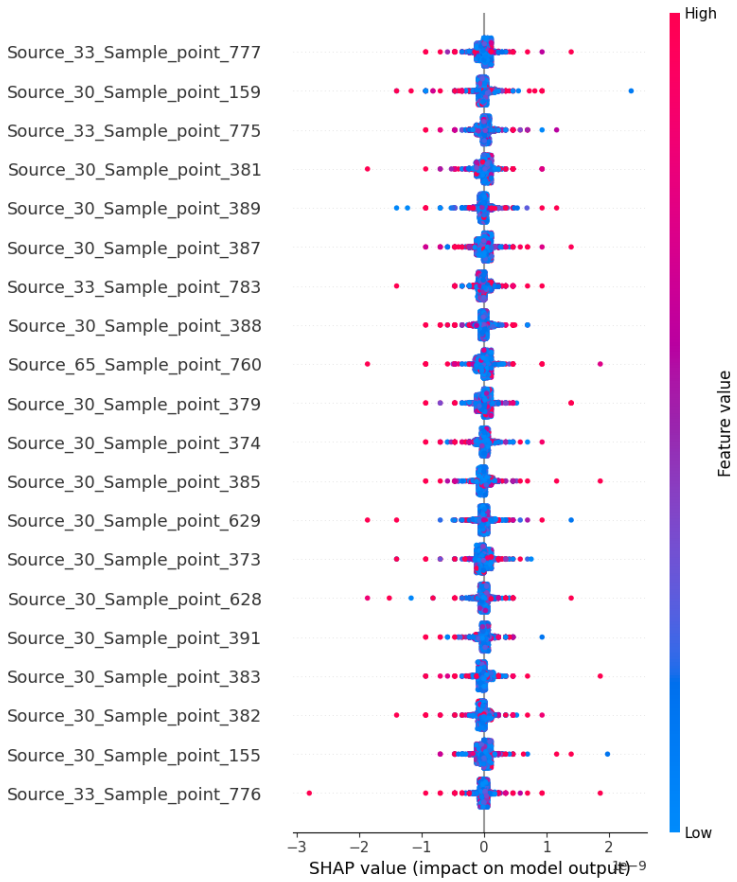


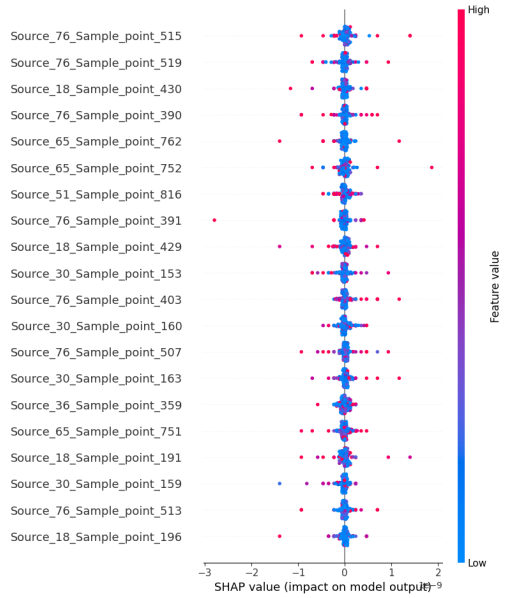
Figure 5.4: Highest 20 SHAP values from all folds

Several insights arise from the analysis on Fig 5.4. First, certain ROIs, like Source_30 (Insula_R) and Source_33 (Cingulum_Mid_L) repeatedly appear among the top 20 SHAP-ranked features, suggesting that these two cortical regions play an really important role in the discrimination of fatigued from rested states. The insula is located in the deep fold between the temporal and frontal lobes, and is widely regarded as a crucial hub for interoceptive processing, emotional regulation, and awareness of bodily states [39]. So the elevated SHAP values in the insular cortex may reflect how shifts in bodily or affective cues help signal the onset of fatigue. Meanwhile, the mid-cingulate cortex, which falls along the dorsal portion of the cingulate gyrus, is implicated in attention allocation, conflict monitoring, and other executive control functions [40]. Because mental fatigue strongly

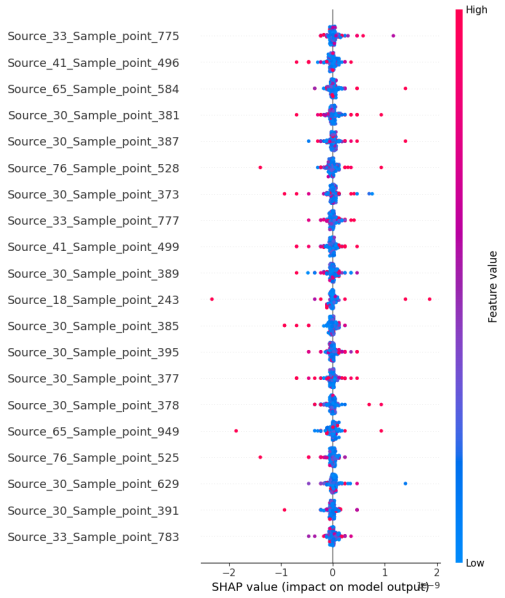
impacts one's ability to maintain cognitive control and handle task demands, it is not surprising that activity in this region would emerge as a key driver of the model's output. Furthermore, multiple distinct timepoints for these same sources appear among the most impactful features, indicating that not only do these regions matter, but the precise timing of their EEG fluctuations is equally critical. Observing these high SHAP values across different epochs shows the CNN-BiLSTM's reliance on how neural dynamics in these sources evolve over time and the patterns that evidently help the network differentiate fatigue states.

These findings in a neurophysiological scope align with the established literature that, for instance, frontal and parietal sources signals often correlate with fatigue. By identifying the features with high SHAP values, we can be sure that the network bases its prediction of fatigue vs rested stated in meaningfull spatio-temporal EEG features. This finding also strengthens the overall robustness of our pipeline which starts from source localization of the EEG signals to identify the cortical source and ends with a CNN-BiLSTM network for classification. Lastly, using SHAP together with source localization provides really good information about why certain features exert strong influences on the model's classification. It demonstrates that the model's internal learned representations coincide with known neural correlates of fatigue, ultimately offering a blueprint that is interpretable for future researchers aiming to refine EEG-based fatigue detection.

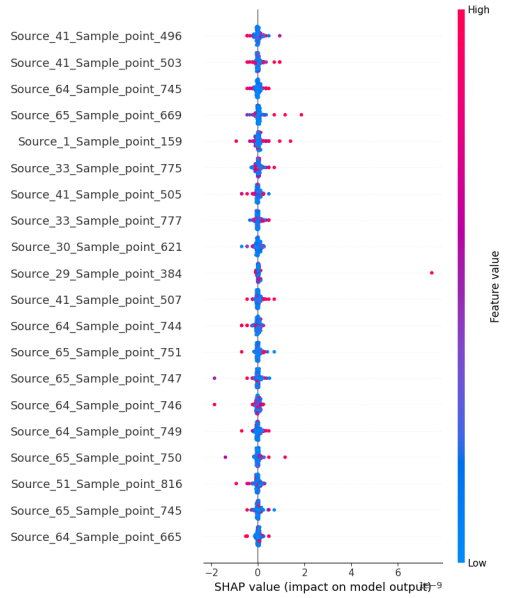
Below we can see in detail the SHAP values on the experiments on each fold:



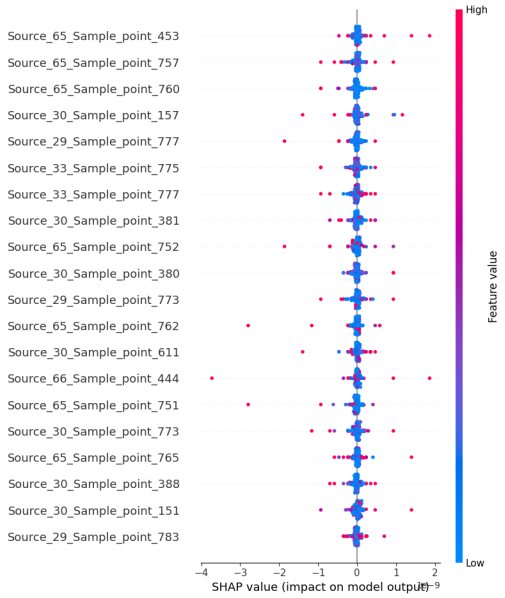
(a) SHAP values after training on the 1st fold



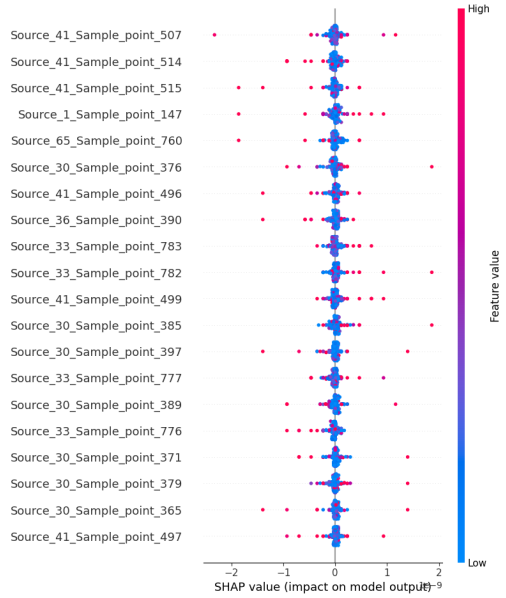
(b) SHAP values after training on the 2nd fold



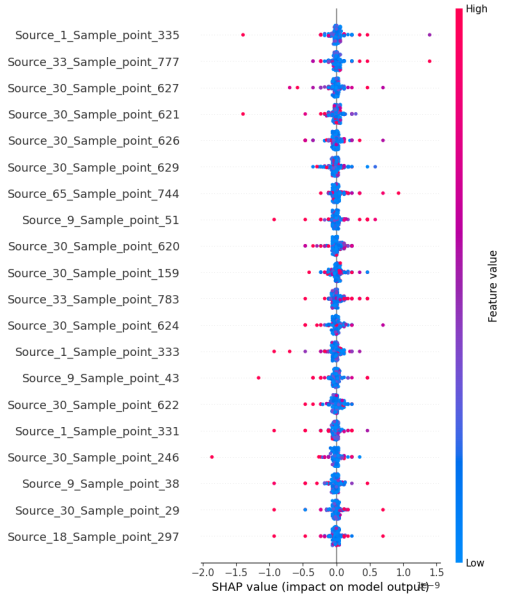
(c) SHAP values after training on the 3rd fold



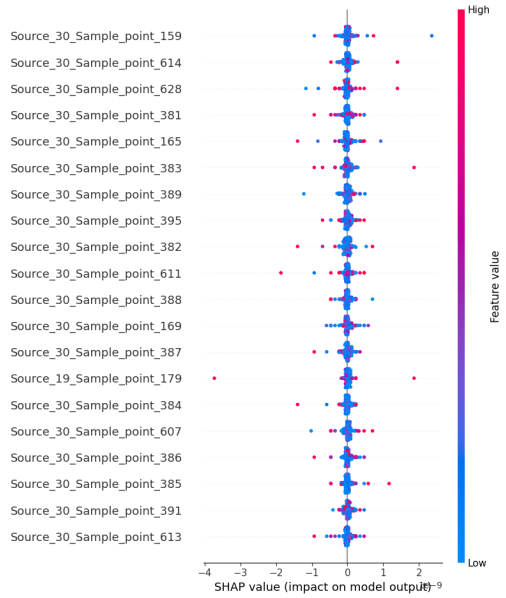
(d) SHAP values after training on the 4th fold



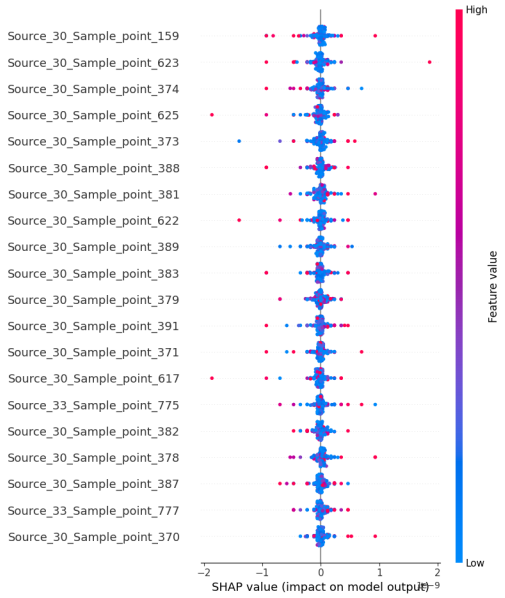
(e) SHAP values after training on the 5th fold



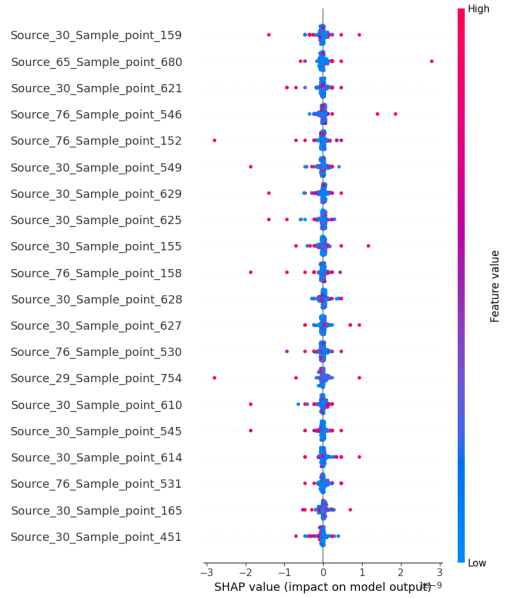
(f) SHAP values after training on the 6th fold



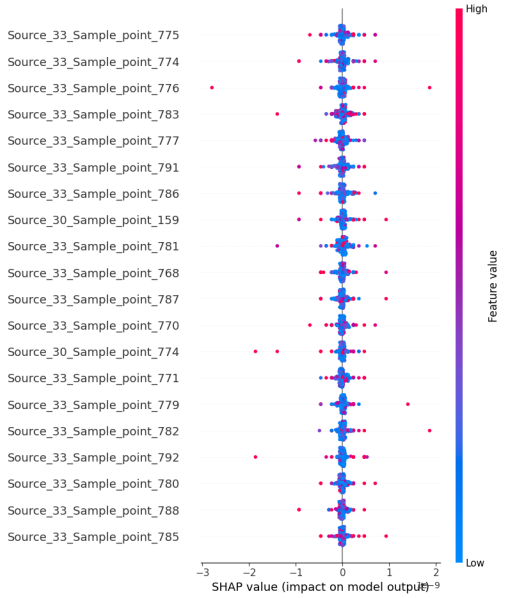
(g) SHAP values after training on the 7th fold



(h) SHAP values after training on the 8th fold



(i) SHAP values after training on the 9th fold



(j) SHAP values after training on the 10th fold

Table 5.2: Top 8 sources identified in SHAP analysis, showing the number of folds in which each source appeared as a significant feature

Source #	**Brain Region (ROI Label)**	**Folds**
30	Insula_R	1, 2, 3, 4, 5, 6, 7, 8, 9, 10
33	Cingulum_Mid_L	2, 3, 4, 5, 6, 8, 10
65	Precuneus_L	1, 2, 3, 4, 5, 6, 9
41	Calcarine_L	2, 3, 5
76	Temporal_Mid_R	1, 2, 9
18	Rolandic_Oper_R	1, 2, 6
29	Insula_L	3, 4, 9
1	Precentral_L	3, 5, 6

By analysing the results of the top SHAP values of every fold in the cross-validation scheme we can access a comprehensive understanding of the features that the CNN-BiLSTM model used in order to classify the mental fatigue states. We can see that some key features consistently contribute to the model's predictions across all folds. These are some critical brain regions and time points that are important for distinguishing between Fatigue and Rest states. Several sources, which represent brain regions (ROIs), like Source_33, and Source_30 frequently appeared as the most impactful across all folds. Source_30 and Source_33 seem to have the most impact as they were the only sources that appeared in the previous highest 20 SHAP values across all folds. This shows that these regions correspond to cortical regions involved in cognitive load and fatigue. As presented in Table 5.2 we can see that other sources also appear frequently in many of the folds on the cross-validation scheme, which shows that there are more brain regions important in mental fatigue assessment.

Firstly we can see that Source 30 which corresponds to the right insular cortex (Insula R) appears in almost all the folds (1 through 10). The right insula is widely recognized for its role in processing interoceptive signals those originating within the body, such as heart rate, respiration, and gut sensations and linking them to emotional or cognitive states [41]. In situations that demand sustained effort, the right insula can become a key mediator of subjective feelings like stress or discomfort. As task difficulty accumulates, an individual might experience heightened arousal or shifts in autonomic markers that the right insula helps interpret. This region also facilitates transitions between focusing on external stimuli and monitoring internal bodily status, a dynamic frequently stressed by fatigue. When mental fatigue sets in, changes in insular activity may be an early physiological alarm. The insula can modulate attention and working-memory processes based on internal feedback, such that an overtaxed state in the body is signaled through increased or dysregulated insular output. Within the SHAP analysis, the repeated appearance of Source_30 in every

fold underscores how robustly the right insula's signals distinguish between rested and fatigued states. This signal consistency suggests that fatigue influences bodily awareness in a reproducible way, making right-insula activation an important marker for classification models.

Source_33 which corresponds to the left mid-cingulate cortex (Cingulum_Mid_L) plays a central role in cognitive control, conflict monitoring, and adaptive behavior. It is integral to detecting performance lapses, such as when a subject becomes prone to missing stimuli or responding incorrectly in cognitively demanding tasks. During extended task performance, neural resources in the MCC can become depleted, leading to slower or less efficient error processing. From a fatigue perspective, when participants are pushed to sustain attention over time, the MCC may exhibit reduced capacity to handle competing demands. This manifests as subtle time-domain EEG fluctuations lower amplitude responses or altered latencies that the CNN-BiLSTM architecture captures. Source_33 shows in most folds (2, 3, 4, 5, 6, 8, 10) and reveals that erratic MCC activity is a strong indicator of entering a fatigued state. In real-world terms, diminished MCC involvement may translate to difficulty maintaining set goals or adapting to quick, unexpected changes in the task environment, both hallmarks of mental fatigue [42].

One source that appears frequently other than the previous dominant sources is Source_65 which is the region left precuneus (Precuneus_L) and is located in the superior parietal region. The left precuneus is associated with a variety of functions, such as visuospatial processing, self-referential thought, and attentional control. Cognitive neuroscientists often regard the precuneus as pivotal for juggling internal mental imagery and externally oriented tasks, making it highly relevant in sustaining working-memory load. In an n-back paradigm, for instance, participants must repeatedly update and compare stimuli, a process that draws heavily on parietal resources [43]. Accordingly, Source_65 appears as a significant marker in folds 1, 2, 3, 4, 5, 6, and 9 which is an indicator that the left precuneus reliably flags the strain of continuous mental workload.

Another important feature is the Source_41 which is located on the left calcarine cortex (Calcarine_L) also known as primary visual cortex. This region, found in the occipital lobe, corresponds to the primary visual cortex, responsible for the initial cortical processing of visual input [44]. Even though mental fatigue research often focuses on frontal and parietal lobes, fatigue can also alter early perceptual mechanisms. When individuals are cognitively drained, top-down attention on visual stimuli may fluctuate, leading to small changes in the amplitude and timing of neural responses in this region. This source appears in three folds (2, 3, 5) showing it is an important brain region for mental fatigue assessment.

The Source_76 which is the right middle temporal gyrus (Temporal_Mid_R) is heavily implicated in semantic processing, language-related comprehension, and the integration of multimodal sensory data. In intensive cognitive tasks, the middle temporal gyrus helps bridge incoming information (visual, auditory, or linguistic) with stored knowledge, enabling rapid decoding of stimuli. Mental fatigue may degrade this efficiency, manifesting as more variable or attenuated EEG signals. This helps explain why Source_76 stands out

in Folds 1, 2, and 9, marking it as one of good temporal-lobe correlates of fatigue [45].

Situated near the junction of somatosensory and motor cortex, the Source_18 which is the rolandic operculum in the right hemisphere (Rolandic_Oper_R) is integral to sensorimotor integration. Seeing Source_18 frequently within folds 1, 2, and 6 underscores that fatigue is not confined to purely cognitive areas rather, it permeates into motor-preparatory mechanisms. Once sensorimotor gating is compromised, participants might exhibit delayed or inconsistent response times, a hallmark behavioral outcome of mental fatigue.

In Folds 3, 4, and 9, mirroring Source_30 (right insula), Source_29 represents the left insula (Insula_L). Both sides contribute to emotional regulation and bodily awareness, although there can be subtle lateralization effects some studies link the left insula more closely with cognitive aspects of emotional processing, whereas the right insula is sometimes more attuned to autonomic or arousal states. Bilateral insular involvement suggests a widespread interoceptive imbalance under fatigue, reflecting the individual's experience of strain or rising stress internally [46].

The left precentral gyrus (Precental_L) which is the Source_1 and is dedicated to voluntary motor control for the right side of the body appears in folds (3, 5, 6). Often considered outside the "core" network of cognition, it can nonetheless be a sensitive index of fatigue where if an individual's motor signals degrade or slow down, it may indicate the broader system's inability to sustain good attentional or executive processes. In the repeated, rapid-response environment of the n-back, small changes in motor-planning readiness can serve as early warning signs of deteriorating cognitive ability [47].

Chapter 6

Discussion

This thesis set out to determine whether a lightweight, CNN-BiLSTM pipeline, trained on source-localized EEG signals, can distinguish a mentally fatigued state from a rested state with high reliability while remaining physiologically interpretable. A 10-fold cross-validation scheme showed an average accuracy of 91.55% and a narrow 2.67% standard-deviation of the values of the accuracies where the network made 1874 correct predictions out of 2047 trials, showing that the network learned task-general features rather than being tied to a specific partition of the data. The confusion matrices reveal that the model detects fatigue slightly more readily than rest, an outcome that is expected when the fatigued condition is accompanied by stronger and more stereotypical oscillatory changes, whereas rest may slip into light mind-wandering or incipient drowsiness that partially mimics fatigue in EEG dynamics.

High performance alone is not sufficient, the central question is whether the network bases its decision on plausible neuro-cognitive cues. To address this, a two-step interpretability analysis was carried out. First, the raw surface EEG was projected to a set of 80 sources, providing an anatomically meaningful representation. Second, SHAP values were computed to identify the spatio-temporal samples that have the greatest influence on the final classification. A really consistent picture emerged across folds where the right insular cortex (Source 30) emerged as the single most influential region, appearing in every fold and frequently contributing SHAP values exceeding 0.12. The insula's role in mental awareness and autonomic regulation makes it a good connection for the subjective sense of exhaustion shown in EEG rhythms. Also the left mid-cingulate (Source 33) repeatedly appeared among the most influential sources, showing in 7 of 10 folds, consistent with its involvement in performance monitoring and conflict detection. Additional regions included the left precuneus (Source 65, 7 folds) and to a lesser degree, the left calcarine cortex (Source 41, 3 folds), right middle-temporal gyrus (Source 76, 3 folds), and bilateral opercular and pre-central regions. This pattern fits with the established understanding that mental fatigue modulates the salience-network hub in the insula (mediating interoceptive awareness and autonomic status), the cingulo-opercular system that sustains performance monitoring, and parietal sites that do working-memory updating. The pres-

ence of primary visual and sensorimotor regions in the explanation maps suggests that fatigue-related changes spread early in the perceptual stream and into motor-preparatory circuits, in line with behavioural findings that reaction times slow as mental resources fade.

The blend of convolutional filters and bidirectional LSTM layers appears critical for these results. The convolutional stage can isolate spatially distributed, time-specific motives that correlate with fatigue, while the BiLSTM stage captures how these motifs accumulate or dissipate over time. In practical terms, the bidirectional component helps the network to evaluate a time point in context, observing both the build-up of fatigue markers and their subsequent evolution resulting in reducing false positives that might arise from remaining artifacts. The convergence of SHAP features across folds further underscores that the network relies on robust and generalisable EEG signatures rather than on chance correlations.

Despite these encouraging findings, several limitations deserve attention. First, the dataset was acquired under a single cognitive-load paradigm, namely the N-back task. While the task is a well-validated stressor, it captures only one side of the complex phenomenon of mental fatigue. Future work could evaluate the network on settings such as extended driving or industrial monitoring, where sources of fatigue are multifactorial and there are many external distractions. Second, although the inverse model and 80 region parcellation simplified the analysis and reduced computational load, higher-resolution source modelling or subject-specific anatomical models could sharpen the localisation of fatigue-critical generators. Third, the experiment adopted a binary labelling scheme. A richer annotation along a continuous fatigue spectrum or at least three or more discrete fatigue tiers could show a better feedback and help bridge the gap between laboratory classifications and the gradual decrease observed in real-world alertness.

Taken together, the study shows that a relatively light neural architecture, supplemented by source localisation and SHAP-based explanation, can provide reliable, interpretable discrimination of mental fatigue from rest. By capturing both spatial patterns and temporal evolution, the model lays a foundation for wearable or in-vehicle system for monitoring fatigue that can operate in real time. Extending the methodology to broader tasks, larger cohorts, and continuous fatigue metrics will move the technology from more theoritical toward practical deployment in safety-critical domains such as aviation, transportation, and high-demand knowledge work.

Chapter 7

Conclusion and future directions

In conclusion, this thesis introduced the CNN-BiLSTM network for EEG-based mental fatigue assessment that can capture both the spatial representations of brain signals with 2D convolutional layers and the temporal dependencies with bidirectional LSTMs. By applying a thorough preprocessing and a 10-fold cross-validation strategy, the model had a final average accuracy of 91.55%, which idndicates its robustness and reliability across multiple partitions of the dataset. Something more crucial than the raw performance was the SHAP analysis that provided insight into why certain features like time points, and brain regions (ROIs) contributed so strongly to distinguishing the fatigued and rested states. This interpretability is really important in order to correlate all the learned features with already known neurophysiological signatures of fatigue.

A key strength of the proposed pipeline is its end-to-end nature, rather than relying only on hand-crafted features, the network learned spatio–temporal patterns directly from multi-channel EEG. By simplifying the model's internal operations to a convolutional front-end followed by BiLSTM components, we ensured that the short-term and long-term EEG dynamics were both leveraged for a good prediction. This design choice proved to be really effective not just for classification accuracy but also for offeringa simple and lightweight architecture compared to more complex or specialized neural architectures in this problems. Moreover, the state-of-the-art interpretability method (SHAP) was integrated in the solution and it helped to clarify that the CNN-BiLSTM latched onto meaningful EEG fluctuations over time, and showing that the detected features are not arbitrary artifacts but relevant cognitive changes related to fatigue.

There are, however, quite a number of ways that this might be taken further scientifically and technologically. For instance, this may extend to more extended and longitudinal EEG sets, possibly with repeated measures over days or weeks, which would reveal just how invariant or time-varying these fatigue signatures are in naturalistic usage. Methodologically, the study could include cross-subject generalization, crucial for establishing that an approach being put forward can perform when trained on some and tested on unseen subjects, this being a pretty tough requirement for any practical deployment across a diverse range of populations. After all, real-time implementation is the next step for a range

of applications like aviation, driving, and industrial workplace settings where continuous monitoring of mental fatigue can forestall all types of accidents. Low-latency inference can only be achieved by compressing models or optimizing them for specific hardware and/or by employing much smaller yet powerful architecture.

One direction is going to be further multimodal integration, so we are going to complement EEG, for example, with eye-tracking data, or facial electromyography, or heart rate variability, where all these biosignals will allow us to include a more holistic measure of fatigue, and thus the model will be able to handle the artifacts in the signal and gaps in the data a lot better. Such fusion will also make the system more robust in case some sensors fail or return noisy signals. In parallel, deeper investigation of the explanation of models beyond SHAP may show if some sampling points or regions of the cortex prevail over time during decision-making, thus yielding more diverse neurophysiological insights. Lastly, we take our findings to clinical and workplace settings, profiting from the collaboration with neuroscientists, ergonomists, and industry partners, allowing iterative refinements of the approach to fit the domain-specific needs and constraints.

Generally, this thesis provides appropriate evidence for how such a combination of CNN and BiLSTM architecture together with the interpretability analyses can successfully classify and explain the mental fatigue states from EEG signals. Further chapters in this study can scale up the methodology to larger scenarios while keeping things light and real-time, furthering insights into the brain-based pattern analysis of cognitive fatigue. This pipeline, under continued emphasis on explainable deep learning, wide curation of data, and practical deployment scenarios, has the possibility to keep improving both scientific understanding and mitigation at the levels of mental fatigue in applied, real-world settings.

Bibliography

[1] Queensland Brain Institute, "Brain anatomy," 2024, accessed: 2024-11-23. [Online]. Available: https://qbi.uq.edu.au/brain/brain-anatomy

[2] Kenhub, "Lobes of the brain: Structure and function," 2023, accessed: 2024-11-28. [Online]. Available: https://www.kenhub.com/en/library/anatomy/lobes-of-the-brain

[3] TESS Research Foundation, "All about neurons," 2021, accessed: 2024-11-28. [Online]. Available: https://www.tessresearch.org/neurons/

[4] A. B. Book, "Functional asl analysis: Task-based fmri," n.d., accessed: 2024-11-28. [Online]. Available: https://andysbrainbook.readthedocs.io/en/stable/ASL/fASL_03_Task.html

[5] S. Geirnaert, "Signal processing algorithms for eeg-based auditory attention decoding," Ph.D. dissertation, May 2022.

[6] M. Richter, M. Kurpas, and M. Maska, "Learning by confusion approach to characterize phase transitions," 06 2022.

[7] D. van der Linden, M. Frese, and T. F. Meijman, "Mental fatigue and the control of cognitive processes: effects on perseveration and planning," *Acta Psychologica (Amst)*, vol. 113, no. 1, pp. 45–65, May 2003.

[8] W. Klimesch, "Eeg alpha and theta oscillations reflect cognitive and memory performance: a review and analysis," *Brain Research Reviews*, vol. 29, no. 2, pp. 169–195, 1999.

[9] A. M. Owen, K. M. McMillan, A. R. Laird, and E. Bullmore, "N-back working memory paradigm: a meta-analysis of normative functional neuroimaging studies," *Human Brain Mapping*, vol. 25, no. 1, pp. 46–59, May 2005.

[10] D. van der Linden, M. Frese, and T. F. Meijman, "Mental fatigue and the control of cognitive processes: effects on perseveration and planning," *Acta Psychologica*, vol. 113, no. 1, pp. 45–65, 2003.

[11] M. A. S. Boksem and M. Tops, "Mental fatigue: Costs and benefits," *Brain Research Reviews*, vol. 59, no. 1, pp. 125–139, 2008.

[12] N. Tiwari, D. R. Edla, S. Dodia, and A. Bablani, "Brain computer interface: A comprehensive survey," *Biologically Inspired Cognitive Architectures*, vol. 26, pp. 118–129, 2018.

[13] G. H. Klem, H. O. Lüders, H. H. Jasper, and C. Elger, "The ten-twenty electrode system of the international federation: The international federation of clinical neurophysiology," *Electroencephalogr Clin Neurophysiol Suppl*, vol. 52, pp. 3–6, 1986.

[14] D. van der Linden, M. Frese, and T. F. Meijman, "Mental fatigue and the control of cognitive processes: effects on perseveration and planning," *Acta Psychologica*, vol. 113, no. 1, pp. 45–65, 2003.

[15] E. Wascher, B. Rasch, J. Sänger, S. Hoffmann, D. Schneider, G. Rinkenauer, H. Heuer, and M. Gutberlet, "Frontal theta activity reflects distinct aspects of mental fatigue," *Biological psychology*, vol. 96, 12 2013.

[16] A. Craik, Y. He, and J. L. Contreras-Vidal, "Deep learning for electroencephalogram (eeg) classification tasks: a review," *Journal of Neural Engineering*, vol. 16, no. 3, p. 031001, apr 2019.

[17] R. T. Schirrmeister, J. T. Springenberg, L. D. J. Fiederer, M. Glasstetter, K. Eggensperger, M. Tangermann, F. Hutter, W. Burgard, and T. Ball, "Deep learning with convolutional neural networks for eeg decoding and visualization," *Human Brain Mapping*, vol. 38, no. 11, pp. 5391–5420, 2017.

[18] D. Das Chakladar, S. Dey, P. P. Roy, and D. P. Dogra, "Eeg-based mental workload estimation using deep blstm-lstm network and evolutionary algorithm," *Biomedical Signal Processing and Control*, vol. 60, p. 101989, 2020.

[19] C. Fan, J. Hu, S. Huang, Y. Peng, and S. Kwong, "Eeg-tnet: An end-to-end brain computer interface framework for mental workload estimation," *Frontiers in Neuroscience*, vol. 16, p. 869522, 2022.

[20] A. Gupta, G. Siddhad, V. Pandey, P. P. Roy, and B.-G. Kim, "Subject-specific cognitive workload classification using eeg-based functional connectivity and deep learning," *Sensors*, vol. 21, no. 20, 2021.

[21] D. D. Chakladar, S. Datta, P. P. Roy, and V. A. Prasad, "Cognitive workload estimation using variational autoencoder and attention-based deep model," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 15, no. 2, pp. 581–590, 2023.

[22] J. Yedukondalu, K. Sunkara, V. Radhika, S. Kondaveeti, M. Anumothu, and Y. M. Krishna, "Cognitive load detection through eeg lead wise feature optimization and ensemble classification," *Scientific Reports*, vol. 15, no. 1, p. 842, 2025.

[23] X. Xu, J. Tang, T. Xu, and M. Lin, "Mental fatigue degree recognition based on relative band power and fuzzy entropy of eeg," *International Journal of Environmental Research and Public Health*, vol. 20, no. 2, p. 1447, Jan 2023.

[24] Z. Xing, E. Dong, J. Tong, Z. Sun, and F. Duan, "Application of mental fatigue classification in cross task paradigm," in *2022 IEEE International Conference on Mechatronics and Automation (ICMA)*, 2022, pp. 1750–1754.

[25] P. Zhang, X. Wang, J. Chen, W. You, and W. Zhang, "Spectral and temporal feature learning with two-stream neural networks for mental workload assessment," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 27, no. 6, pp. 1149–1159, 2019.

[26] M. Su, W. Li, F. Peng, W. Zhou, R. Zhang, and Y. Wen, "Eeg-based mental fatigue detection using cnn-lstm," in *2022 16th ICME International Conference on Complex Medical Engineering (CME)*, 2022, pp. 302–305.

[27] G. Siddhad, P. P. Roy, and B.-G. Kim, "Neural networks meet neural activity: Utilizing eeg for mental workload estimation," in *Pattern Recognition*, A. Antonacopoulos, S. Chaudhuri, R. Chellappa, C.-L. Liu, S. Bhattacharya, and U. Pal, Eds. Cham: Springer Nature Switzerland, 2025, pp. 325–339.

[28] T. Hartley and G. J. Hitch, "Working memory," 10 2022. [Online]. Available: https://oxfordre.com/psychology/view/10.1093/acrefore/9780190236557.001.0001/acrefore-9780190236557-e-768

[29] T. S. Redick and D. R. B. Lindsey, "Complex span and n-back measures of working memory: A meta-analysis," *Psychonomic Bulletin & Review*, vol. 20, no. 6, pp. 1102–1113, 2013.

[30] S. Makeig, A. Bell, T.-P. Jung, and T. J. Sejnowski, "Independent component analysis of electroencephalographic data," in *Advances in Neural Information Processing Systems*, D. Touretzky, M. Mozer, and M. Hasselmo, Eds., vol. 8. MIT Press, 1995.

[31] W. A. Mir, M. Anjum, Izharuddin, and S. Shahab, "Deep-eeg: An optimized and robust framework and method for eeg-based diagnosis of epileptic seizure," *Diagnostics*, vol. 13, no. 4, 2023.

[32] A. M. Dale and M. I. Sereno, "Improved localizadon of cortical activity by combining eeg and meg with mri cortical surface reconstruction: A linear approach," *Journal of Cognitive Neuroscience*, vol. 5, no. 2, pp. 162–176, 04 1993.

[33] R. Grech, T. Cassar, J. Muscat, K. P. Camilleri, S. G. Fabri, M. Zervakis, P. Xanthopoulos, V. Sakkalis, and B. Vanrumste, "Review on solving the inverse problem in eeg source analysis," *Journal of NeuroEngineering and Rehabilitation*, vol. 5, no. 1, p. 25, Nov 2008.

[34] R. D. Pascual-Marqui, "Standardized low-resolution brain electromagnetic tomography (sloreta): Technical details," *Methods and Findings in Experimental and Clinical Pharmacology*, vol. 24 Suppl D, pp. 5–12, 2002.

[35] C. M. Michel and D. Brunet, "Eeg source imaging: A practical review of the analysis steps," *Frontiers in Neurology*, vol. Volume 10 - 2019, 2019.

[36] "On the effects of data normalization for domain adaptation on eeg data," *Engineering Applications of Artificial Intelligence*, vol. 123, p. 106205, 2023.

[37] H. Gupta and S. Kalla, "Analysis of micro state eeg signals through rnn classification," in *2021 Innovations in Power and Advanced Computing Technologies (i-PACT)*, 2021, pp. 1–6.

[38] E. Karim, H. R. Pavel, A. Jaiswal, M. Zadeh, M. Theofanidis, G. Wylie, and F. Makedon, "An eeg-based cognitive fatigue detection system," 08 2023, pp. 131–136.

[39] A. D. B. Craig, "How do you feel — now? the anterior insula and human awareness," *Nature Reviews Neuroscience*, vol. 10, no. 1, pp. 59–70, 2009.

[40] G. Bush, P. Luu, and M. I. Posner, "Cognitive and emotional influences in anterior cingulate cortex," *Trends in Cognitive Sciences*, vol. 4, no. 6, pp. 215–222, 2000.

[41] H. D. Critchley, S. Wiens, P. Rotshtein, A. Öhman, and R. J. Dolan, "Neural systems supporting interoceptive awareness," *Nature Neuroscience*, vol. 7, no. 2, pp. 189–195, 2004.

[42] C. B. Holroyd, J. J. F. Ribas-Fernandes, D. Shahnazian, M. Silvetti, and T. Verguts, "Human midcingulate cortex encodes distributed representations of task progress," *Proceedings of the National Academy of Sciences*, vol. 115, no. 25, pp. 6398–6403, 2018.

[43] A. E. Cavanna and M. R. Trimble, "The precuneus: A review of its functional anatomy and behavioural correlates," *Brain: A Journal of Neurology*, vol. 129, no. Pt 3, pp. 564–583, 2006.

[44] D. Ress, J. T. Heeger, E. P. Simoncelli, R. E. Engel, and D. J. Tootell, "Activity in primary visual cortex predicts performance in a visual detection task," *Nature Neuroscience*, vol. 3, no. 9, pp. 940–945, 2000.

[45] G. Darnai, A. Matuz, H. A. Alhour, G. Perlaki, G. Orsi, Ákos Arató, A. Szente, E. Áfra, S. A. Nagy, J. Janszky, and Árpád Csathó, "The neural correlates of mental fatigue and reward processing: A task-based fmri study," *NeuroImage*, vol. 265, p. 119812, 2023.

[46] V. Menon and L. Q. Uddin, "Saliency, switching, attention and control: A network model of insula function," *Brain Structure and Function*, vol. 214, no. 5-6, pp. 655–667, 2010.

[47] H. D. Critchley, S. Wiens, P. Rotshtein, A. Öhman, and R. J. Dolan, "Neural systems supporting interoceptive awareness," *Nature Neuroscience*, vol. 7, no. 2, pp. 189–195, 2004.

Appendix A

Python Code

```
import scipy.io
import glob
import pandas as pd
import numpy as np
import tensorflow as tf
import matplotlib.pyplot as plt
import itertools
import shap
from tensorflow import keras
from tensorflow.keras import activations
from tensorflow.keras.utils import to_categorical
from tensorflow.keras import layers
from tensorflow.keras.models import Sequential
from tensorflow.keras.backend import is_keras_tensor
from sklearn.model_selection import train_test_split
from sklearn.model_selection import KFold
from sklearn.preprocessing import StandardScaler
from sklearn.decomposition import PCA
from tensorflow.keras.optimizers import Adam
from tensorflow.keras.callbacks import EarlyStopping
from sklearn.metrics import confusion_matrix
from sklearn import preprocessing
from sklearn.metrics import f1_score
from numba import cuda
import gc

gpus = tf.config.list_physical_devices('GPU')
if gpus:
  # allocate specific size of memory on the GPU
  try:
    tf.config.set_logical_device_configuration(
        gpus[0],
        [tf.config.LogicalDeviceConfiguration(memory_limit=13336)])
```

```
    logical_gpus = tf.config.list_logical_devices('GPU')
    print(len(gpus), "Physical GPUs,", len(logical_gpus), "Logical GPUs")
  except RuntimeError as e:
    print(e)

batch = 32
epochs = 50
channels = 80
sampling_points = 1024
class_names = ['Fatigue', 'Rest']

mat_data = []
labels = []

files = glob.glob(r'C:\Users\giannos\Desktop\data\Fatigue_asc\*.mat')

# Loop through and load each file with 4 second segmentation
# different segmentation lengths were tested with optimal 4 seconds
for i in range(2):
    if i>0:
        files = glob.glob(r'C:\Users\giannos\Desktop\data\Rest_asc\*.mat')
    for file in files:
        data = scipy.io.loadmat(file)
        d = np.array(data['EEG_source'][:1024])
        mat_data.append(d)
        labels.append(i)
        # mat_data.append(d[256:512])
        # labels.append(i)
        # mat_data.append(d[512:768])
        # labels.append(i)
        # mat_data.append(d[768:1024])
        # labels.append(i)


features=np.array(mat_data)
labels=np.array(labels)

features = features.reshape([len(features), channels, sampling_points, 1])
print(features.shape)
print(labels.shape)

# CNN-BiLSTM model
def cnn_model():

    model = Sequential()
    inp = layers.Input(shape=(channels, sampling_points, 1))
```

```
80
81      conv1 = layers.Conv2D(32, kernel_size=(3, 3), activation='tanh', padding='
            same', input_shape=(channels, sampling_points, 1))(inp)
82      #x22 = layers.Dropout(0.5)(x1)
83      conv1 = layers.MaxPooling2D((2, 2), padding='valid')(conv1)
84      conv2 = layers.Conv2D(64, ( 5, 5), activation='tanh', padding='same')(conv1)
85      conv2 = layers.MaxPooling2D((2, 2), padding='valid')(conv2)
86      # x444 = layers.Dropout(0.8)(conv2)
87      conv3 = layers.Conv2D(128, (7, 7), activation='tanh', padding='same')(conv2)
88      conv3 = layers.MaxPooling2D((2, 2), padding='valid')(conv3)
89
90      x13 = layers.Dropout(0.3)(conv3)
91      x14 = layers.TimeDistributed(layers.Flatten())(x13)
92      x14 = layers.TimeDistributed(layers.Dense(128, activation='relu'))(x14)
93      x14 = layers.Bidirectional(layers.LSTM(128, return_sequences=True))(x14)
94
95      x14 = layers.Bidirectional(layers.LSTM(128))(x14)
96      #with tf.device("cpu:0"):
97
98      #x14 = layers.Dropout(0.5)(x14)
99      x15 = layers.Dense(1024, 'relu')(x14)
100     x16 = layers.Dropout(0.3)(x15)
101     x16 = layers.Dense(128, 'relu')(x16)
102
103     out = layers.Dense(2, 'softmax')(x16)
104     model = tf.keras.Model(inputs=inp, outputs=out)
105     model.summary()
106     # Compile the model
107     model.compile(optimizer=Adam(learning_rate=0.0001, decay=0.00001), loss='
            categorical_crossentropy', metrics=['accuracy'])
108     return model
109
110
111 pat = 10 # this is the number of epochs with no improvment after which the
        training will stop
112 early_stopping = EarlyStopping(monitor='val_loss', patience=pat, verbose=1)
113
114 all_acc = []
115 all_shap_values = []
116 all_val_samples = []
117
118
119 def fit_and_evaluate(t_x, val_x, t_y, val_y, fold_num, EPOCHS=epochs, BATCH_SIZE=
        batch):
120     model = None
121     model = cnn_model()
```

```
122     history = model.fit(t_x, t_y, epochs=EPOCHS, batch_size=BATCH_SIZE, callbacks
            =[early_stopping],
123                         verbose=1, validation_data=[val_x,val_y])
124     _, acc_t = model.evaluate(t_x, t_y)
125     print('training accuracy:', str(round(acc_t * 100, 2)) + '%')
126     _, acc = model.evaluate(val_x, val_y)
127     all_acc.append(acc)
128     print('testing accuracy:', str(round(acc * 100, 2)) + '%')
129     # print("Val Score: ", model.evaluate(val_x, val_y))
130
131
132     # SHAP Explainer
133     #use the whole train set as background
134     background = t_x
135     explainer = shap.GradientExplainer(model, background)
136
137     #use the whole val set
138     val_sample = val_x
139     print(f'Validation sample shape: {val_sample.shape}')
140
141
142     shap_values = explainer.shap_values(val_sample)
143     print(f'SHAP values shape (before sum/flatten): {np.array(shap_values).shape}
            ')
144
145
146     shap_values = np.sum(shap_values, axis=-1) # remove the class dimension
147
148     # flatten SHAP values
149     shap_values_flat = shap_values.reshape(val_sample.shape[0], -1)
150     val_sample_flat = val_sample.reshape(val_sample.shape[0], -1)
151
152     print(f'SHAP values shape (after flatten): {shap_values_flat.shape}')
153     print(f'Validation sample shape (after flatten): {val_sample_flat.shape}')
154
155     all_shap_values.append(shap_values_flat)
156     all_val_samples.append(val_sample_flat)
157
158
159     assert shap_values_flat.shape[1] == val_sample_flat.shape[1], \
160         f"Mismatch in shape: SHAP values {shap_values_flat.shape[1]}, validation
                sample {val_sample_flat.shape[1]}"
161
162
163     feature_names = [f'Source_{i}_Sample_point_{j}' for i in range(channels) for
            j in range(sampling_points)]
```

```
164
165     # create SHAP plot
166     shap.summary_plot(shap_values_flat, val_sample_flat, feature_names=
            feature_names, show=False)
167     plt.savefig(f'C:/Users/giannos/Desktop/python/shap_fold_{fold_num}.png') #
            Save SHAP plot
168     plt.close()
169     return history,model
170
171
172
173 n_folds = 10
174
175 # save the model history in a list after fitting so that we can plot later
176 model_history = []
177 predicted_targets = np.array([])
178 actual_targets = np.array([])
179 all_shap = []
180
181 kfold = KFold(n_splits=n_folds, shuffle=True, random_state=1)
182
183 i = 1
184 for train_index, val_index in kfold.split(features, labels):
185
186     y = to_categorical(labels, num_classes=2, dtype="int32")
187
188     print("Training on Fold: ", i)
189
190     X_train, X_val = features[train_index], features[val_index]
191     y_train, y_val = y[train_index], y[val_index]
192     print(len(X_train))
193     scaler = None
194     scaler = StandardScaler()
195     x_train = scaler.fit_transform(X_train.reshape(len(X_train) * sampling_points
            , channels)).reshape(X_train.shape)
196     x_val = scaler.transform(X_val.reshape(len(X_val) * sampling_points ,
            channels)).reshape(X_val.shape)
197
198     print(x_train.shape)
199     #print(x_test.shape)
200     print(x_val.shape)
201     hist,mod = fit_and_evaluate(x_train, x_val, y_train, y_val, i, epochs, batch)
202     model_history.append(hist)
203
204
205     predicted_labels = mod.predict(x_val)
```

```
206     predicted_labels = np.where(predicted_labels>0.5 , 1, 0)
207     predicted_targets = np.append(predicted_targets, tf.argmax(predicted_labels,
            axis=1))
208     actual_targets = np.append(actual_targets, tf.argmax(y_val, axis=1))
209 
210     #mod.save('C:/Users/giannos/Desktop/seed_preprocessed/models/subject'+str(
            subj)+'/model'+str(subj)+'_'+str(i)+'.h5')
211     i=i+1
212     # device = cuda.select_device(0)
213     # device.reset()
214     gc.collect()
215 
216 def plot_confusion_matrix(predicted_labels_list, y_val_list):
217     cnf_matrix = confusion_matrix(y_val_list, predicted_labels_list)
218     np.set_printoptions(precision=2)
219 
220 # plot non-normalized confusion matrix
221     plt.figure()
222     generate_confusion_matrix(cnf_matrix, classes=class_names, title='Confusion
            matrix, without normalization')
223     plt.savefig(r'C:\Users\giannos\Desktop\confusion1.png')
224     #plt.show()
225 
226 # plot normalized confusion matrix
227     plt.figure()
228     generate_confusion_matrix(cnf_matrix, classes=class_names, normalize=True,
            title='Normalized confusion matrix')
229     plt.savefig(r'C:\Users\giannos\Desktop\confusion2.png')
230     #plt.show()
231 
232 def generate_confusion_matrix(cnf_matrix, classes, normalize=False, title='
    Confusion matrix'):
233     if normalize:
234         cnf_matrix = cnf_matrix.astype('float') / cnf_matrix.sum(axis=1)[:, np.
                newaxis]
235         print("Normalized confusion matrix")
236     else:
237         print('Confusion matrix, without normalization')
238 
239     plt.imshow(cnf_matrix, interpolation='nearest', cmap=plt.get_cmap('Blues'))
240     plt.title(title)
241     plt.colorbar()
242 
243     tick_marks = np.arange(len(classes))
244     plt.xticks(tick_marks, classes, rotation=45)
245     plt.yticks(tick_marks, classes)
```

```
246
247     fmt = '.2f' if normalize else 'd'
248     thresh = cnf_matrix.max() / 2.
249
250     for i, j in itertools.product(range(cnf_matrix.shape[0]), range(cnf_matrix.
            shape[1])):
251         plt.text(j, i, format(cnf_matrix[i, j], fmt), horizontalalignment="center"
                ,
252                  color="white" if cnf_matrix[i, j] > thresh else "black")
253
254     plt.tight_layout()
255     plt.ylabel('True label')
256     plt.xlabel('Predicted label')
257
258     return cnf_matrix
259
260 # after collecting all SHAP values and validation samples from all folds
        concatenate them
261 all_shap_values = np.concatenate(all_shap_values, axis=0)
262 all_val_samples = np.concatenate(all_val_samples, axis=0)
263
264 # Number of top features to show
265 top_n = 20 # you can adjust this value to show more or fewer top features
266
267 #calculate the mean absolute SHAP values across all samples for each feature
268 mean_abs_shap_values = np.mean(np.abs(all_shap_values), axis=0)
269
270 # get the indices of the top N features based on the mean absolute SHAP values
271 top_n_indices = np.argsort(mean_abs_shap_values)[-top_n:]
272
273 # filter SHAP values and corresponding validation samples to keep only the top N
274 top_shap_values = all_shap_values[:, top_n_indices]
275 top_val_samples = all_val_samples[:, top_n_indices]
276
277 #generate feature names
278 feature_names = [f'Source_{i}_Sample_point_{j}' for i in range(channels) for j in
        range(sampling_points)]
279
280 # filtered feature names to keep only the top_n
281 top_feature_names = [feature_names[i] for i in top_n_indices]
282
283 shap.summary_plot(top_shap_values, top_val_samples, feature_names=
        top_feature_names, show=False)
284
285 plt.savefig(r'C:\Users\giannos\Desktop\python\shap_all_folds.png')
286 plt.close()
```

```
287 #print(predicted_targets)
288 plot_confusion_matrix(predicted_targets, actual_targets)
289
290 print("Average accuracy : "+ str(round(np.mean(all_acc) * 100, 2)) + "% , std : "
        + str(round(np.std(all_acc) * 100, 2)) + "%")
```