Multimodal Deep Learning for Driver Monitoring: Integrating EEG and Vision for Robust Drowsiness Detection and Safety Enhancement

Shams Ur Rahman

B.Sc. Computer Science

Supervised by Prof Noel O'Connor and Dr Graham Healy

A thesis presented for the degree of Doctor of Philosophy

School of Computing

Dublin City University

September 2025

Declaration

I hereby certify that this material, which I now submit for assessment on the programme of study leading to the award of Doctor of Philosophy is entirely my own work, and that I have exercised reasonable care to ensure that the work is original, and does not to the best of my knowledge breach any law of copyright, and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the text of my work.

Signed: Shams Ur Rahman ID No: 21260130

Date: 01/09/2025

Dedication

To my beloved wife, Dr. Taranum Naz, whose love, patience and belief in me carried me through the toughest days of this PhD. To my little daughter, Zimal, who was born as I was finishing the final months of this journey and filled my life with joy and purpose. To my mom, whose prayers, guidance and unconditional support have always been my foundation. And to my dad and all my family members, for their heartfelt wishes and prayers that gave me strength along the way.

List of Publications

The research presented in this thesis has led to the following peer-reviewed publications and submissions:

Journal Articles

1. **Rahman, S. U.**, O'Connor, N. E., Lemley, J., & Healy, G. (2025). An investigation of pre-stimulus EEG for prediction of driver reaction time. *Biomedical Physics & Engineering Express*. (Accepted).
   (Forms the basis of parts of Chapter 4 of this thesis, focusing on EEG parameter optimization and 1D-CNN for reaction time prediction.)

2. **Rahman, S. U.**, O'Connor, N. E., Lemley, J., Parsi, A., & Healy, G. (2025). Predicting Driver Drowsiness Using Multimodal Transformers on EEG and Vision Data. *IEEE Transactions on Intelligent Vehicles*. (Submitted).
   (Forms the basis of Chapter 6 of this thesis, detailing the end-to-end multimodal transformer for drowsiness detection.)

3. **Rahman, S. U.**, O'Connor, N. E., & Healy, G. (2025). Enhancing Driver Reaction Time Prediction via Vision-Based EEG Analysis. *IEEE Access*. (Submitted).
   (Forms the basis of Chapter 5 of this thesis, detailing the application of ResNet18 and Vision Transformers to image-transformed EEG for reaction time prediction.)

Conference and Workshop Papers

4. **Rahman, S. U.**, O'Connor, N. E., Lemley, J., & Healy, G. (2022). Using Prestimulus EEG to Predict Driver Reaction Time To Road Events. In *Proceedings of the 44th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (pp. 4036-4039). IEEE. (Accepted and Published).
   (Forms the basis of initial parts of Chapter 4 of this thesis, establishing the feasibility of EEG-based reaction time prediction with classical machine learning.)

5. **Rahman, S. U.**, O'Connor, N. E., & Healy, G. (2025). Efficient Transformer-Based Drowsiness Detection on the Edge using a Hybrid MobileViT-LSTM Architecture. In *ACM Workshop on Automotive and Medical Multimedia: Bridging the Gap Between Mobility and Healthcare*. (Accepted).
   (Forms the basis of Chapter 7 of this thesis, detailing the MobileViT-LSTM model for edge-based drowsiness detection.)

Contents

List of Figures

List of Tables

Multimodal Deep Learning for Driver Monitoring: Integrating EEG and Vision for Robust Drowsiness Detection

Shams Ur Rahman

Abstract

Road accidents remain a major global concern, with driver drowsiness and delayed reaction times recognized as key contributing factors. This thesis advances driver-monitoring research by developing multimodal approaches that integrate electroencephalography (EEG) and vision data to predict reaction times and drowsiness.

The investigation first demonstrates that pre-stimulus EEG signals—specifically spectral power in the alpha and theta bands—contain rich information for estimating reaction times to critical road events. Using subject-independent machine-learning pipelines, short EEG windows recorded before event onset effectively differentiate between fast and slow responses. The work then explores the benefits of incorporating vision data as a second modality by fusing EEG signals with camera-based observations of the driver. One branch converts EEG power-spectral-density features into image-like representations for analysis with deep convolutional neural networks and transformer models. Another branch directly integrates raw EEG signals with synchronised video frames through end-to-end multimodal transformer architectures. Results indicate that transformers equipped with cross-modal attention capture complex interdependencies between neural and visual cues, yielding significant improvements in driver-drowsiness detection over unimodal approaches.

Real-time deployment is addressed by designing and optimising a lightweight pipeline for edge-based processing. This resource-efficient model enables rapid analysis of facial cues under diverse driving conditions, ensuring operation on embedded platforms such as

smartphones and automotive edge devices.

Extensive evaluations on large-scale simulated datasets confirm the generalisability of the proposed approaches across varied driving scenarios. Experiments reveal that transformer-based fusion significantly enhances predictive performance by effectively combining complementary neural and visual cues. Moreover, the lightweight pipeline maintains high accuracy under stringent computational constraints, enabling real-time, on-device deployment.

Chapter 1

Introduction

1.1 The Imperative for Advanced Driver State Monitoring

The ubiquity of vehicular transport is a defining feature of modern society, underpinning economic vitality and individual mobility. However, this convenience is shadowed by the persistent and grave risk of road traffic accidents. These incidents exact a devastating global toll in terms of human lives lost and injuries sustained, concurrently imposing substantial economic and societal burdens [1]. A significant and preventable factor contributing to a large proportion of these accidents is the compromised state of the driver, particularly impairments arising from fatigue, drowsiness, and lapses in sustained attention [2, 3, 4]. When a driver's cognitive faculties are diminished, their capacity for accurate hazard perception, sound decision-making, and rapid motor responses is critically impaired, often leading to catastrophic outcomes [5, 6]. Consequently, the rigorous development of robust, reliable, and real-time systems for monitoring and interpreting driver state transcends mere academic inquiry; it represents a critical imperative for substantially enhancing road safety and mitigating the incidence of preventable vehicular tragedies.

Traditional methodologies for assessing driver impairment, such as subjective self-report questionnaires [7] or performance-based laboratory vigilance tasks [8], while valu-

able within controlled research paradigms, are largely unsuited for continuous, non-intrusive monitoring within the complex and dynamic milieu of an operating vehicle. This inherent limitation has catalyzed extensive research efforts focused on leveraging physiological and behavioural signals. These efforts, coupled with sophisticated advancements in signal processing and machine learning, aim to forge intelligent Advanced Driver Assistance Systems (ADAS) endowed with the capability to detect, and ideally preempt, the emergence of dangerous driver states.

This imperative is not only academic but is now a major focus of the automotive industry and regulatory bodies. The current state-of-the-art in commercially available driver-monitoring systems, found in vehicles from leading manufacturers, primarily relies on vision-based solutions using infrared cameras to track head pose, eye gaze, and blink patterns. While effective, these systems are largely reactive—designed to detect overt signs of impairment rather than predict them—and manufacturers do not publish a single, comparable "accuracy". Independent assessments such as Euro NCAP use scenario-based scoring rather than percentage accuracy. The adoption of these technologies is being accelerated by regulation: the EU's General Safety Regulation requires that all newly registered vehicles from 7 July 2024 be equipped with a Driver Drowsiness and Attention Warning (DDAW) system [9]. The Commission Delegated Regulation (EU) 2021/1341 further specifies validation and performance requirements, including that a warning shall be issued at a drowsiness level equivalent to KSS ≥ 8 (and may be issued at KSS 7) [10]. Furthermore, safety-rating bodies such as Euro NCAP award higher scores to vehicles that incorporate robust driver-monitoring functions in their Safety Assist – Safe Driving protocol, creating a strong market incentive for implementation [11]. This regulatory and commercial landscape underscores the urgent need for the kind of advanced, accurate, and reliable monitoring techniques investigated in this thesis.

The Karolinska Sleepiness Scale (KSS) referenced in these regulations is a standard 9-point Likert-type scale used to measure subjective sleepiness, ranging from 1 ("extremely alert") to 9 ("very sleepy, great effort to keep awake") [12]. Its adoption in regulatory frameworks as a benchmark for drowsiness levels highlights the importance of developing

objective monitoring systems, like those explored in this thesis, that can accurately infer such states without relying on subjective self-reporting.

1.2 Leveraging Physiological and Behavioural Signals for Driver State Assessment

Among the diverse array of physiological signals investigated, Electroencephalography (EEG) has attracted considerable scientific attention. This is primarily due to its unique capacity to provide a direct, non-invasive measurement of cortical brain activity, offering a window into the dynamic neurophysiological processes that underpin cognitive function [13]. Observable fluctuations in EEG patterns, especially within canonical frequency bands such as Alpha (8-12 Hz) and Theta (4-8 Hz), are well-documented correlates of shifts in alertness, variations in cognitive workload, and the insidious progression towards drowsiness and fatigue [14, 15]. This inherent sensitivity makes EEG a prime candidate for objectively assessing the neural basis of driver performance and vigilance.

Concurrently, vision-based monitoring systems have emerged as a highly promising and practical alternative. These systems typically employ cameras to analyze facial video data, extracting behavioural cues indicative of a driver's state. Commonly monitored indicators include the percentage of eye closure (PERCLOS), blink rate and duration, frequency of yawning, and changes in head pose or posture [16, 17]. The increasing ubiquity of high-quality cameras in modern vehicles and consumer smartphones further enhances the appeal and accessibility of vision-based approaches for continuous driver monitoring.

However, neither EEG nor vision-based systems, when utilized in isolation, represent a panacea. EEG recordings can be susceptible to motion artefacts and electrical noise, and traditionally require the application of scalp electrodes, which may raise concerns about user comfort and practicality for everyday use. Vision-based systems, on the other hand, can be adversely affected by fluctuating environmental conditions such as variable lighting or direct glare, and may suffer from occlusions (e.g., from eyewear or

hand-to-face gestures). Moreover, vision systems primarily capture overt behavioural manifestations of impairment, potentially missing the subtle, early-onset neurocognitive changes that precede obvious behavioural signs. This recognition of unimodal limitations has increasingly motivated research into multimodal approaches. Such strategies aim to synergistically integrate complementary information from diverse sensing modalities—like EEG and vision—to achieve a more comprehensive, robust, and accurate assessment of driver state [18, 19].

1.3 Research Scope and Thesis Contributions

This thesis presents a systematic and progressive investigation into the development, application, and evaluation of advanced machine learning and deep learning methodologies for driver state assessment. The research encompasses two primary application domains: the prediction of driver reaction time from pre-event EEG signals and the classification of driver drowsiness using unimodal (EEG, vision) and sophisticated multimodal (EEG-vision fusion) techniques. A significant thematic thread woven throughout this work is the methodical advancement from foundational feasibility studies and model optimization to the exploration of cutting-edge architectures, culminating in addressing the critical practical challenge of deploying accurate and efficient monitoring systems in resource-constrained edge computing environments.

The research detailed herein is designed to make several distinct and impactful contributions to the field:

1. **Establishing and Optimizing EEG-Based Prediction of Driver Reaction Time:** This work rigorously examines the capacity of pre-stimulus EEG spectral features to predict a driver's impending reaction time to critical road events. This includes a systematic exploration of optimal EEG input parameters (e.g., window length, frequency bands, channel selection) and a comparative evaluation of classical machine learning algorithms against specialized 1D Convolutional Neural Networks (1D-CNNs) designed for sequential spectral data.

2. **Advancing EEG Analysis through Vision Model Application to Spectral Images:** A novel approach is investigated where EEG spectral features are transformed into 2D image representations—specifically, PSD Matrix Images and Scalp Topographies. The efficacy of applying standard (ResNet18) and advanced (Vision Transformer, ViT-B/16) deep learning vision models to these EEG-derived images for reaction time prediction is assessed, with the ViT-B/16 demonstrating state-of-the-art performance that surpasses even specialized 1D-CNNs.

3. **Pioneering Advanced Multimodal Fusion for Robust Drowsiness Detection:** The thesis explores the synergistic potential of combining EEG and facial vision data for drowsiness classification. A hierarchy of multimodal fusion strategies is developed and evaluated, progressing from simple feature concatenation to sophisticated transformer-based feature-level fusion, and ultimately culminating in a novel end-to-end multimodal transformer architecture that jointly processes raw EEG and vision data to achieve superior classification accuracy.

4. **Enabling Efficient and Deployable Edge-Based Drowsiness Detection:** Recognizing the critical need for practical, real-world solutions, this research proposes and validates a hybrid efficient vision transformer-LSTM model (MobileViT-LSTM). This model is specifically designed for real-time, vision-based drowsiness detection on resource-constrained mobile devices, demonstrating a viable pathway for deploying advanced AI in critical driver safety applications.

These contributions are pursued through a series of interconnected empirical chapters, each designed to address specific, well-defined research questions. A consistent emphasis is placed on rigorous subject-independent validation methodologies to ensure that the findings are generalizable and hold relevance for diverse driver populations.

1.4 Guiding Research Themes and Questions

The investigative journey of this thesis is structured around three overarching research themes. Each theme addresses a distinct aspect of driver state monitoring, and the

empirical studies in the subsequent chapters are guided by a set of specific research questions aligned with these themes. This structure reflects a progressive deepening of inquiry, from foundational explorations of neural signals to the practical challenges of deploying AI-driven safety systems.

Theme 1: Deciphering Pre-Stimulus Neural Signatures for Anticipatory Performance Prediction

A core ambition in proactive safety research is the development of systems capable of anticipating decrements in driver performance before these decrements manifest as observable errors. This theme delves into the capacity of neural signals, specifically EEG, recorded in the moments immediately preceding a critical event, to predict subsequent behavioural outcomes. The key research questions under this theme, addressed comprehensively in Chapters 4 and 5, are:

- **RQ1: Can pre-stimulus EEG spectral features reliably predict driver reaction time (RT) in a subject-independent framework, and what are the optimal input parameters (e.g., time window, frequency bands, channel subsets) for this task?** (Chapter 4)

- **RQ2: Can a specialized 1D-CNN, designed for sequential data, enhance the predictive performance for RT compared to classical machine learning models?** (Chapter 4)

- **RQ3: Can transforming 1D EEG spectral features into 2D image representations and applying advanced Vision Transformers (ViTs) lead to a new state-of-the-art in RT prediction accuracy, surpassing even specialized 1D models?** (Chapter 5)

Theme 2: Achieving Robust Drowsiness Detection through Synergistic Multimodal Fusion

While unimodal systems provide valuable insights, human drowsiness is a multifaceted phenomenon. This theme is predicated on the hypothesis that by intelligently integrating information from multiple modalities, specifically EEG and facial vision—it is possible to develop drowsiness detection systems that are more robust and accurate than those relying on any single data stream. The central research questions guiding the investigations in Chapter 6 are:

- **RQ1: What are the baseline performances of unimodal EEG (EEGNet) and vision (ResNet18, ViT-Base) models for subject-independent drowsiness classification?**
- **RQ2: Are simple or feature-level fusion strategies sufficient to improve classification accuracy beyond the best-performing unimodal baseline?**
- **RQ3: Can an end-to-end multimodal transformer architecture, by jointly learning from raw EEG and vision data, achieve a synergistic performance uplift and significantly outperform all unimodal and feature-level fusion approaches?**

Theme 3: Translating Advanced AI into Practical, Efficient Edge-Based Solutions

The ultimate societal impact of driver state monitoring systems is contingent upon their practical deployability in real-world environments, which often necessitate operation on resource-constrained edge devices. This third research theme confronts the challenge of translating high-performance deep learning models into efficient and deployable solutions. The investigations in Chapter 7, focusing on the highly informative vision modality, address the following research questions:

- **RQ1: Can a hybrid deep learning architecture (MobileViT-LSTM) that**

combines an efficient vision transformer with a recurrent neural network achieve an optimal balance between high accuracy and computational efficiency for video-based drowsiness detection?

- **RQ2: Is the proposed hybrid model capable of real-time inference on a representative mobile edge device, and can it be successfully exported to a standardized format (ONNX) to facilitate its seamless deployment in practical applications?**

1.5 Thesis Outline

This thesis is meticulously structured to systematically investigate the research themes and questions articulated above, guiding the reader through a coherent progression of studies:

- **Chapter 1 (Current Chapter): Introduction** serves to provide the overarching motivation for the research, delineate its scope and principal contributions, introduce the guiding research themes and their constituent questions, and furnish an outline of the entire thesis structure.
- **Chapter 2: Literature Review** offers a comprehensive survey and critical analysis of the existing body of research pertinent to physiological (primarily EEG-based) and vision-based methodologies for driver state monitoring. This includes a review of common signal processing techniques, machine learning and deep learning models applied to EEG and visual data, established multimodal fusion strategies, and an overview of the challenges inherent in developing and deploying such AI-driven systems.
- **Chapter 3: Datasets and Core Methodologies** provides a detailed exposition of the two primary datasets utilized throughout the empirical investigations: the publicly available Cao et al. [20] EEG dataset, which underpins the reaction time prediction studies; and the internally collected Tobii multimodal dataset, which

forms the basis for the drowsiness classification experiments. This chapter also offers a foundational overview of common signal processing techniques, feature extraction methods, and machine learning principles recurrently employed.

- **Chapter 4: Electroencephalography-Based Prediction of Driver Reaction Time using Pre-Stimulus Neural Activity** constitutes the first empirical study, focusing on Theme 1. It investigates the fundamental feasibility of predicting driver RT from pre-stimulus EEG spectral features, explores the optimization of input parameters (such as pre-stimulus window length, frequency band selection, and EEG channel subsets), and introduces a specialized 1D-CNN architecture to enhance feature learning from 1D spectral data.

- **Chapter 5: Enhancing EEG-Based Reaction Time Prediction through Advanced Vision Model Analysis of Spectral Images** continues the exploration under Theme 1. It investigates an innovative approach where EEG spectral features are transformed into 2D image representations (PSD Matrix Images and Scalp Topographies). The chapter then evaluates the efficacy of applying standard (ResNet18) and advanced (Vision Transformer, ViT-B/16) deep learning vision models to these EEG-derived images for RT prediction, comparing their performance against the 1D-CNN benchmark.

- **Chapter 6: Multimodal Transformer-Based Fusion of EEG and Vision for Driver Drowsiness Detection** addresses Theme 2, transitioning the focus to driver drowsiness classification using the Tobii multimodal dataset. This chapter systematically evaluates unimodal performance baselines (EEGNet, ResNet18, ViT-Base), investigates various feature-level fusion techniques, and culminates in the development and validation of a novel end-to-end multimodal transformer architecture that jointly processes raw EEG and facial video data to achieve state-of-the-art classification accuracy.

- **Chapter 7: Efficient Transformer-Based Drowsiness Detection on Edge Devices using a Hybrid MobileViT-LSTM Architecture** directly tackles

Theme 3, focusing on the critical challenge of practical deployment. It proposes and validates the MobileViT-LSTM hybrid model for efficient, real-time, vision-based drowsiness detection, including rigorous assessments of its inference capabilities on a representative mobile edge device and its successful export to the ONNX format for enhanced portability.

- **Chapter 8: General Discussion, Conclusions, and Future Work** serves to synthesize the key findings from all preceding empirical chapters. It discusses their collective implications in the broader context of driver state monitoring, artificial intelligence, and road safety. This chapter also acknowledges the overall limitations of the conducted research and proposes promising and impactful avenues for future investigation in this rapidly evolving field.

This structured progression is intended to provide the reader with a clear and logical path through the research, building from fundamental explorations of EEG predictivity to the development of advanced multimodal systems and, ultimately, to considerations for practical, efficient, and deployable real-world solutions aimed at enhancing driver safety through intelligent monitoring technologies.

Chapter 2

Literature Review: Physiological and Vision-Based Approaches to Driver State Monitoring

2.1 Introduction: The Challenge of Driver Impairment

Mental fatigue and drowsiness represent significant contributors to road accidents globally [2], often stemming from factors such as sleep deprivation or prolonged focus on monotonous tasks like driving [3]. A direct consequence of driver drowsiness is often a slowed reaction time to critical events encountered on the road [5], which can tragically lead to fatal accidents. This issue extends beyond driving; mental fatigue can result in unsafe practices and diminished performance in occupations demanding sustained operator attention, such as crane operation [21]. The dangers associated with drowsy driving are profound, with studies suggesting it can be as hazardous as driving under the influence of alcohol, as both conditions impair reaction time, attention, and decision-making capabilities [22, 23, 6]. According to the National Highway Traffic Safety Administration (NHTSA), drowsy driving was implicated in approximately 2.6% of all fatal motor vehicle crashes in the United States, resulting in an estimated 846 fatalities [24, 25].

Consequently, the development of reliable methods for assessing and predicting driver impairment, particularly drowsiness and reaction time, is paramount for enhancing road safety.

2.2 Approaches to Assessing Driver State

Various methodologies have been employed to measure or infer a driver's level of alertness or fatigue. These can be broadly categorized into subjective assessments, behavioral tests, physiological measurements, and vision-based monitoring.

2.2.1 Subjective and Behavioral Measures

Mental fatigue or drowsiness can be quantified using methods such as psychometric questionnaires [7] or vigilance tests like the Psycho-motor Vigilance Test (PVT) [8]. For instance, the Karolinska Sleepiness Scale (KSS) [12], a 9-point Likert scale, is commonly used to gather subjective ratings of sleepiness.

2.2.2 Physiological and Vision-Based Sensing

To overcome the limitations of subjective and behavioral tests, passive sensing techniques are generally preferred for real-time driver monitoring. These include measuring physiological signals such as heart rate variability [26, 27, 28, 29] or electrodermal activity (EDA), and ocular metrics like changes in pupil size (pupillography) [30] or electrooculogram (EOG) signals [31, 32]. Behavioral metrics derived from driving patterns, like steering wheel movements [33], have also been explored.

Among the physiological measures, Electroencephalography (EEG) stands out as it directly measures neuro-physiological activity. This makes EEG potentially a more reliable method for obtaining objective measures of fatigue and, critically, the moment-to-moment variations in brain activity that may correlate with reaction times to road events. The link between mental fatigue and specific EEG features has been established in several studies [34, 35, 36].

Alternatively, vision-based systems utilizing cameras offer a non-intrusive approach. These systems analyze facial and ocular features, such as eye closure duration, blink rate [37], head pose, yawning frequency [17, 38], and facial expressions [39], to infer drowsiness [40, 16, 41].

This review will delve deeper into EEG and vision-based methods, as they form the core modalities investigated in this thesis, followed by a discussion on multimodal approaches that combine their strengths.

2.3 EEG-Based Driver State Monitoring

EEG signals provide a rich source of information about brain activity and have been extensively studied for monitoring cognitive states, including attention, fatigue, and drowsiness [42, 43, 44].

2.3.1 EEG Signals and Drowsiness/Attention

Research has consistently highlighted the importance of specific frequency bands within the EEG signal in relation to mental states. The alpha (typically 8-12 Hz) and theta (typically 4-8 Hz) bands are particularly relevant for assessing drowsiness and attention [45]. Studies have shown that under conditions requiring high attentional demand, an increase in theta power alongside a decrease in alpha power is often observed [46]. Conversely, research indicates that an increase in lower alpha power occurs when subjects actively try to remain awake despite feeling sleepy [14]. When sleep is permitted, a decrease in alpha and an increase in theta power is typically observed [14]. Klimesch et al. [14] noted the alpha band as particularly reliable for studying mental state. Further supporting this, Aakerstedt et al. [47], studying industrial workers, found an increase in alpha power minutes before sleep onset, while theta activity increased during sleep itself. Increases in theta and alpha band power are generally associated with decreased alertness and the progression towards drowsiness [13, 14, 15, 48]. These frequency bands, along with others like delta (0.5–4 Hz), beta (12–30 Hz), and gamma (30–100 Hz) [49],

represent distinct characteristics of ongoing EEG activity [50].

2.3.2 EEG for Reaction Time Prediction

While much research focuses on classifying drowsiness or mental fatigue [51, 52, 53, 54], the prediction of driver reaction time (RT) using EEG has received comparatively less attention.

Some studies have explored the relationship between EEG features and reaction time. For instance, Foong et al. [52] investigated the use of dry frontal EEG electrodes to predict driver reaction time, using a 2-minute window centered around event onset. They found a positive correlation between RT and delta band power, and negative correlations with other bands.

Other related research has explored driver reaction times in simulated and real-world driving environments. Jurecki et al. [55] studied the relationship between Time-to-Collision (TTC) and driver RT in a real-world experiment involving mock pedestrians, confirming a linear relationship and measuring RTs for braking, steering, and accelerator operation under varying TTC conditions.

2.3.3 Machine Learning Techniques for EEG Analysis

Various machine learning (ML) and feature engineering strategies have been applied to EEG data for driver state assessment. Early work often involved extracting features, such as Power Spectral Density (PSD) computed using methods like Welch's [56, 57], from specific frequency bands (delta, theta, alpha, beta) within defined time windows. These features were then fed into classical ML algorithms. For instance, Liu et al. [54] evaluated logistic regression for mental fatigue recognition.

Deep learning (DL) models have gained traction due to their ability to automatically learn hierarchical features from complex data like EEG [58, 59]. Convolutional Neural Networks (CNNs) are commonly used. Lawhern et al. [60] proposed EEGNet, a compact CNN architecture specifically designed for EEG-based brain-computer interfaces (BCIs) and classification tasks, utilizing depthwise and separable convolutions. Cui et al. [51]

used a CNN model for drowsiness detection, implementing a Global Pooling Layer to learn local features via class activation maps. Their model reportedly learned significant features like alpha spindles and theta bursts, achieving higher accuracy (73.22%) than conventional ML techniques. In another study, Cui et al. [61] again used a CNN, identifying similar features and achieving 78.35% accuracy. Research continues to explore various deep learning architectures, including 1D-CNNs, for processing EEG time-series or derived features, aiming to capture complex temporal and spectral patterns relevant to cognitive states.

2.3.4 Subject-Dependent vs. Subject-Independent Models

A critical consideration in developing practical EEG-based systems is the distinction between subject-dependent and subject-independent models [62]. Subject-dependent models are trained and tested using data from the same individual, potentially capturing subject-specific EEG patterns but limiting generalizability. Many early studies adopted this approach. For real-world deployment, where pre-calibration for every user is impractical, subject-independent models are necessary. These models are trained on data from a group of subjects and tested on entirely new, unseen subjects. Achieving good performance with subject-independent models is challenging due to significant inter-subject variability in EEG signals [63, 64, 65]. Techniques like transfer learning have been explored to bridge this gap. For example, Liu et al. [54] found that a transfer learning-enabled classifier outperformed logistic regression and EEGNet for mental fatigue recognition. Liu et al. [53] used inter-subject transfer-based learning (Maximum Independent Domain Adaptation - MIDA, and Transfer Component Analysis - TCA) to detect mental fatigue, achieving accuracies around 73% and 68% respectively, also exploring single-channel prediction using Random Forest for channel selection.

2.3.5 Investigating EEG Parameters for Optimal Performance

Optimizing the EEG analysis pipeline involves careful consideration of several parameters.

- **Time Window Selection:** The duration of the EEG segment used for analysis is

crucial [49]. Selecting an appropriate window length involves balancing the need to capture sufficient information against potential noise or variability introduced by longer segments. Studies like [66] suggest that activity closer to a stimulus onset might hold more predictive power.

- **Frequency Band Combinations:** While individual bands such as Alpha and Theta are informative, researchers have explored arithmetic combinations—such as *Theta + Beta / Alpha* or *Theta / Beta*—which are frequently cited in cognitive performance research [67, 68, 5, 14, 69, 70]. Studies have investigated combinations like alpha/beta and gamma oscillations during attention tasks, suggesting distinct roles (e.g., prestimulus alpha for top-down control, beta/gamma for bottom-up processing) [71, 72]. The coupling of alpha and theta has also been highlighted as critical for attention [73]. However, combining bands might also compound noise [74] or increase signal complexity [75], potentially requiring more sophisticated models.

- **Channel Selection and Spatial Information:** EEG is typically recorded from multiple channels (electrodes) placed according to standardized systems like the 10-20 system [76]. Different brain regions, and thus electrode locations, are associated with different cognitive functions: frontal with decision-making/attention [77, 78], central with motor processes [79], temporal with auditory/language processing [80], occipital with visual processing [81, 82], and parietal with spatial processing/attention [83, 84]. While using all available channels is common, research explores whether subsets of channels are sufficient or even advantageous [85]. Understanding the spatial distribution of relevant EEG activity is important for interpretation [86]. Techniques like Common Spatial Patterns (CSP) [87] can be used to find spatial filters that maximize discriminability between conditions (e.g., fast vs. slow RT) based on signal variance/power, providing insights into which electrode locations contribute most. However, interpreting these patterns can be complex as they may represent combined activity from multiple sources [88], potentially requiring source localization techniques for deeper understanding [89].

2.4 Vision-Based Driver Drowsiness Detection

Vision-based systems offer a non-intrusive alternative or complement to physiological sensors for monitoring driver state.

2.4.1 Visual Cues for Drowsiness

These systems typically rely on cameras monitoring the driver's face and analyzing various visual indicators associated with fatigue. Commonly used cues include ocular measures like Percentage of Eye Closure (PERCLOS), blink frequency and duration [37, 16], and eye gaze. Facial expressions, particularly yawning frequency and duration [17, 38], are also strong indicators. Head pose, such as nodding frequency or sustained downward gaze, provides further evidence of drowsiness [41, 39]. Handcrafted geometric features derived from facial landmarks, like the Eye Aspect Ratio (EAR) and Mouth Aspect Ratio (MAR) [90, 91], were used in earlier systems but often struggled with variability in real-world conditions [92, 93].

2.4.2 Deep Learning for Vision Analysis

Modern vision-based drowsiness detection predominantly uses deep learning, particularly CNNs, to automatically extract relevant features from facial images or video frames [94, 95, 96]. Models like ResNet [97] have been successfully applied, often pre-trained on large image datasets and fine-tuned for the specific task of classifying alertness states based on visual input. CNNs excel at capturing local patterns and textures [98, 99], contributing to improved robustness compared to handcrafted features [100].

More recently, Vision Transformers (ViTs) [101] have emerged as powerful alternatives or complements to CNNs. ViTs utilize self-attention mechanisms, allowing them to model long-range dependencies and capture global spatial relationships within an image more effectively than the inherently local receptive fields of CNNs [102]. This capability is potentially advantageous for detecting subtle, holistic facial cues associated with drowsiness. However, standard ViTs are computationally intensive [103, 104].

2.4.3 Challenges in Vision-Based Methods

Despite their non-intrusiveness, vision-based systems face challenges. Their performance can be significantly affected by variations in environmental conditions, such as poor or changing illumination (e.g., day vs. night driving, shadows). Occlusions, where the driver's face is partially obscured (e.g., by sunglasses, hands, or masks), can hinder feature extraction. Varying camera angles and distances also impact performance. Furthermore, visual cues primarily reflect the physical manifestations of drowsiness; they may not capture the underlying cognitive state as directly as physiological measures like EEG.

2.5 Multimodal Approaches for Enhanced Detection

Given the complementary strengths and weaknesses of individual modalities like EEG and vision, multimodal approaches that combine information from multiple sources have gained significant interest.

2.5.1 Rationale for Multimodality

The core idea behind multimodal fusion is to leverage the unique information provided by each sensor type to achieve more robust and accurate detection than possible with any single modality alone [105]. For example, EEG provides direct insight into neural activity related to alertness [13, 106] but can be sensitive to noise and inter-subject variability [63]. Vision offers non-intrusive monitoring of behavioral cues [16] but is susceptible to environmental factors and occlusions [17]. By combining them, the system can potentially exploit EEG's sensitivity to internal state changes and vision's ability to capture overt behavioral signs, mitigating the limitations of each [18, 19]. Previous studies have demonstrated the benefits of multimodal fusion in related areas like emotion recognition [19], cognitive workload assessment [107], and fatigue detection [18]. Lian et al. [18] combined EEG and eye-tracking data, using a cross-modal predictive alignment module to improve fusion efficiency. Zhang et al. [108] fused EEG and EOG signals, achieving improved accuracy for drowsiness detection.

2.5.2 Fusion Strategies

Various strategies exist for fusing data from multiple modalities. A common approach is feature-level fusion, where features are first extracted independently from each modality (e.g., using unimodal networks), and the resulting feature vectors are then combined, often by simple concatenation. These fused features are then fed into a classifier (e.g., Bayesian Ridge Classification [109], SVM, ANN). While straightforward, this approach may not fully capture complex inter-modal dependencies, and its effectiveness can be limited if the chosen features or the fusion classifier are suboptimal.

An alternative is decision-level fusion, where each modality is processed by a separate classifier, and their individual predictions are combined (e.g., through voting or weighted averaging).

More advanced strategies involve intermediate or hybrid fusion, and increasingly, end-to-end learning approaches. End-to-end models aim to learn both feature representations and the fusion process jointly, directly from the raw input data of multiple modalities. This allows the model greater flexibility to discover optimal representations and cross-modal interactions without being constrained by pre-defined feature extraction steps.

2.5.3 Transformers in Multimodal Fusion

The success of transformer architectures in natural language processing [110] and computer vision [111] has spurred their application in multimodal tasks. Transformers' core mechanism, self-attention (and its extension, cross-attention), is well-suited for modeling dependencies both within and between different data streams. Multimodal Transformers [112] have been proposed for tasks like sentiment analysis and emotion recognition. Attention mechanisms have also been used with biosignals, for instance, in EEG-based emotion recognition [113].

For fusing EEG and vision, transformer models can be designed to operate at the feature level, applying cross-modal attention to features from unimodal encoders, or in an end-to-end fashion, processing raw data streams directly with transformer-based encoders and integrating them via attention mechanisms. Investigating the effectiveness of such

advanced fusion architectures for driver drowsiness remains an active area of research.

2.6 Real-Time Deployment on Edge Devices

While developing accurate models is crucial, a practical driver monitoring system must operate in real-time on resource-constrained platforms typically found in vehicles or mobile devices (edge computing). This presents significant computational challenges.

2.6.1 The Edge Computing Challenge

Advanced deep learning models, particularly those involving transformers like ViT, often have large parameter counts and high computational demands (FLOPs), making them difficult to deploy on devices with limited processing power, memory, and energy budgets [103]. Achieving low latency inference, essential for timely warnings in safety-critical applications like drowsiness detection, is a primary hurdle.

2.6.2 Efficient Model Architectures

Addressing the deployment challenge requires designing or adapting models for efficiency. Lightweight CNN architectures have been explored, but recent efforts focus on creating efficient transformer variants. MobileViT [114] is a notable example, designed to combine the strengths of CNNs (local feature extraction, inductive biases) with the global context modeling capabilities of transformers, while significantly reducing parameter count and computational cost compared to standard ViTs. Such models aim to bridge the gap between performance and efficiency for vision tasks on mobile platforms.

2.6.3 Temporal Modeling for Edge Vision

Drowsiness is an inherently temporal phenomenon; its onset and progression occur over time [115, 116]. Simply applying an efficient spatial feature extractor like MobileViT on a frame-by-frame basis may miss crucial temporal dynamics. Therefore, for video-based analysis on the edge, architectures need to incorporate temporal modeling. Hybrid

approaches combining efficient spatial feature extractors (like MobileViT or lightweight CNNs) with recurrent neural networks (RNNs), particularly Long Short-Term Memory (LSTM) networks [117], are a strategy explored in literature. LSTMs are adept at capturing temporal dependencies in sequential data [118, 119, 120], potentially allowing a model to aggregate information over a time window to make a more informed prediction.

2.6.4 Optimization Techniques for Deployment

Beyond architectural choices, several optimization techniques are vital for edge deployment. Mixed-precision training (e.g., using Automatic Mixed Precision - AMP [121]) can significantly reduce memory footprint and potentially speed up inference by using lower-precision numerical formats (like float16) for computations where possible, without substantial loss in accuracy. Model quantization [122, 123] further reduces model size and computational cost by representing weights and activations with fewer bits (e.g., 8-bit integers). Pruning techniques remove redundant parameters from the model. Finally, exporting the trained model to standardized formats like ONNX (Open Neural Network Exchange) [124] facilitates deployment across various hardware platforms using optimized inference runtimes (e.g., ONNX Runtime, TensorFlow Lite) that leverage hardware acceleration capabilities [125].

2.7 Summary and Research Gaps

The literature highlights a clear need for effective driver drowsiness and reaction time monitoring systems to improve road safety. While various subjective, behavioral, physiological (EEG, EOG, ECG, etc.), and visual methods exist, EEG and vision-based approaches, particularly using deep learning, have shown significant promise.

EEG analysis, focusing on bands like alpha and theta, can provide direct insights into neural correlates of fatigue and attention. Machine learning, evolving from classical methods with feature engineering (PSD) to deep learning (CNNs, EEGNet, 1D-CNNs), has enabled increasingly sophisticated analysis. A key challenge remains the development

of robust subject-independent models. Research has explored optimizing EEG parameters like time windows, band combinations, and channel configurations.

Vision-based methods offer non-intrusive monitoring of facial cues using CNNs and, more recently, ViTs. However, they face challenges related to environmental variability and may not capture internal states as directly as EEG.

Multimodal approaches, combining modalities like EEG and vision, aim to overcome individual limitations and provide more robust detection. Fusion strategies range from simple feature-level concatenation to potentially more advanced end-to-end models employing attention mechanisms.

Finally, the practical deployment of these systems necessitates addressing the computational constraints of edge devices. This involves developing efficient model architectures (e.g., MobileViT), incorporating temporal modeling for video analysis (e.g., using LSTMs), and employing optimization techniques like mixed-precision training and standardized deployment formats (ONNX).

This review identifies several areas where further investigation is warranted and which motivate the research presented in this thesis:

- Exploring the feasibility and optimization of predicting driver reaction time using pre-stimulus EEG signals with robust subject-independent models.

- A deeper investigation into advanced multimodal fusion techniques, particularly leveraging transformer architectures, to effectively combine EEG and vision data for enhanced drowsiness detection compared to unimodal and simpler fusion methods.

- Addressing the critical gap between high-performance deep learning models and the requirements of real-time edge deployment, focusing on developing and optimizing architectures that are both accurate and computationally efficient for vision-based monitoring on resource-constrained platforms.

Addressing these aspects aims to advance the state-of-the-art towards more accurate, reliable, and deployable driver state monitoring systems.

The drowsiness classification studies presented in this thesis will leverage a large-scale multimodal dataset where 'Alert' and 'Drowsy' states are robustly defined by a controlled sleep-deprivation protocol and validated by subjective Karolinska Sleepiness Scale (KSS) scores, allowing for a clear and unambiguous ground truth for model training and evaluation.

Figure 2.1 provides a high-level visualization of the structure and flow of the literature reviewed in this chapter. The diagram begins by outlining the core challenge of driver impairment (Section 2.1), encompassing issues like drowsiness and slowed reaction times, which motivates the exploration of various assessment strategies (Section 2.2). It then delves into the specifics of the primary monitoring approaches discussed: EEG-based methods (Section 2.3) and Vision-based methods (Section 2.4), highlighting their respective inputs, techniques, and inherent challenges. The diagram shows how the limitations and complementary nature of these individual modalities motivate the investigation into Multimodal Fusion techniques (Section 2.5), particularly those leveraging advanced architectures like transformers. Moving towards practical application, the figure illustrates the critical considerations for Edge Deployment (Section 2.6), covering both the challenges and potential solutions like efficient models (e.g., MobileViT) and optimization techniques. Finally, the diagram culminates in the specific Research Gaps and Thesis Focus (Section 2.7), demonstrating how the reviewed literature informs and positions the contributions of this work in optimizing EEG analysis, advancing multimodal fusion, and enabling efficient edge models.

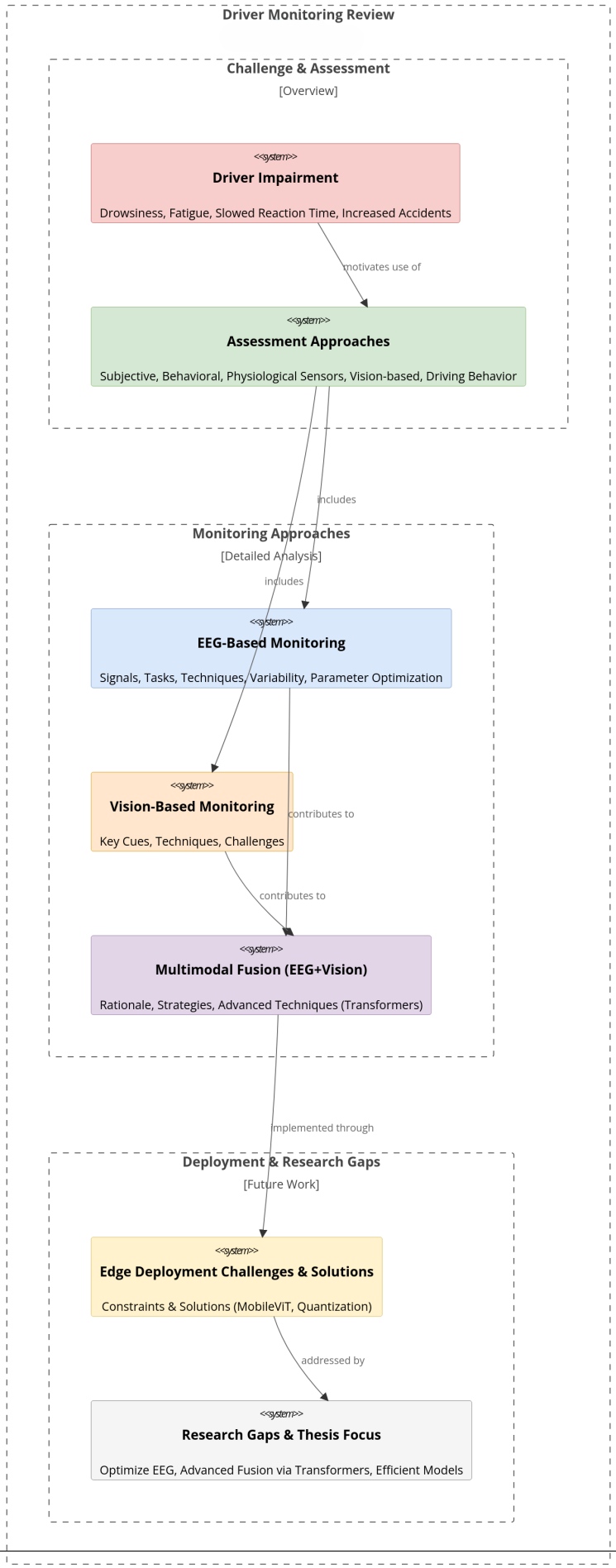


Figure 2.1: Literature Review: Driver State Monitoring

Chapter 3

Datasets and Core Methodologies

3.1 Introduction

This chapter serves as a foundational element for the experimental work presented in this thesis. It provides detailed descriptions of the two primary datasets utilized for developing and evaluating models related to driver state assessment: a publicly available multi-channel Electroencephalography (EEG) dataset focused on sustained attention and reaction time, and a comprehensive, internally collected multimodal dataset capturing various physiological and behavioural signals during induced drowsiness.

Furthermore, this chapter offers a high-level overview of the core technical concepts and methodologies explored in the subsequent empirical chapters (Chapters 4 through 7). It introduces the progression of techniques applied, ranging from classical machine learning on EEG features to advanced deep learning architectures, including Convolutional Neural Networks (CNNs) and Transformers, applied to both unimodal and multimodal data, culminating in models optimized for edge deployment. Finally, it outlines the consistent evaluation framework adopted throughout the thesis, emphasizing the importance of subject-independent validation for ensuring the generalizability of the findings in real-world driver monitoring contexts. This chapter aims to equip the reader with the necessary background on the data and the key analytical approaches employed, setting the stage for the detailed investigations presented later.

3.2 Multi-Channel EEG Driving Dataset (Cao et al.)

The first dataset employed extensively in Chapters 4 and 5 is a publicly available collection focused on capturing driver behaviour and brain dynamics during a sustained-attention driving task, as described by Cao et al. [20]. This dataset is crucial for investigating the neural correlates of reaction time (RT) fluctuations in response to driving events.

3.2.1 Experimental Design and Task

The experiment involved 27 voluntary participants (students or staff, aged 22-28 years) operating an immersive driving simulator equipped with a six-degree-of-freedom motion platform. Participants were instructed to maintain their simulated vehicle in the center of a four-lane highway during a monotonous, 90-minute night-time driving scenario designed to induce fatigue and vigilance decrements. Crucially, the simulation incorporated an event-related lane-departure paradigm. Randomly, the simulated vehicle would begin to drift either left or right from the cruising lane (deviation onset). Participants were required to quickly correct this deviation by steering the vehicle back to the center of the original lane (response onset to response offset). Reaction Time (RT) for each event was defined as the duration between the deviation onset and the response onset. The experiment was designed to isolate steering responses, with participants not required to control acceleration or braking. Figure 3.1 represent the visual representation of the driving setup and the event markers [20].

3.2.2 Data Acquisition

For each of the 62 sessions recorded across the 27 participants, the following data were acquired simultaneously:

- **EEG Data:** Continuous 32-channel EEG signals were recorded using an Ag/AgCl cap placed according to a modified international 10-20 system, referenced to linked mastoids. Data were acquired using a Scan SynAmps2 Express system (Compumedics Ltd.) at a sampling rate of 500 Hz with 16-bit resolution. Impedance

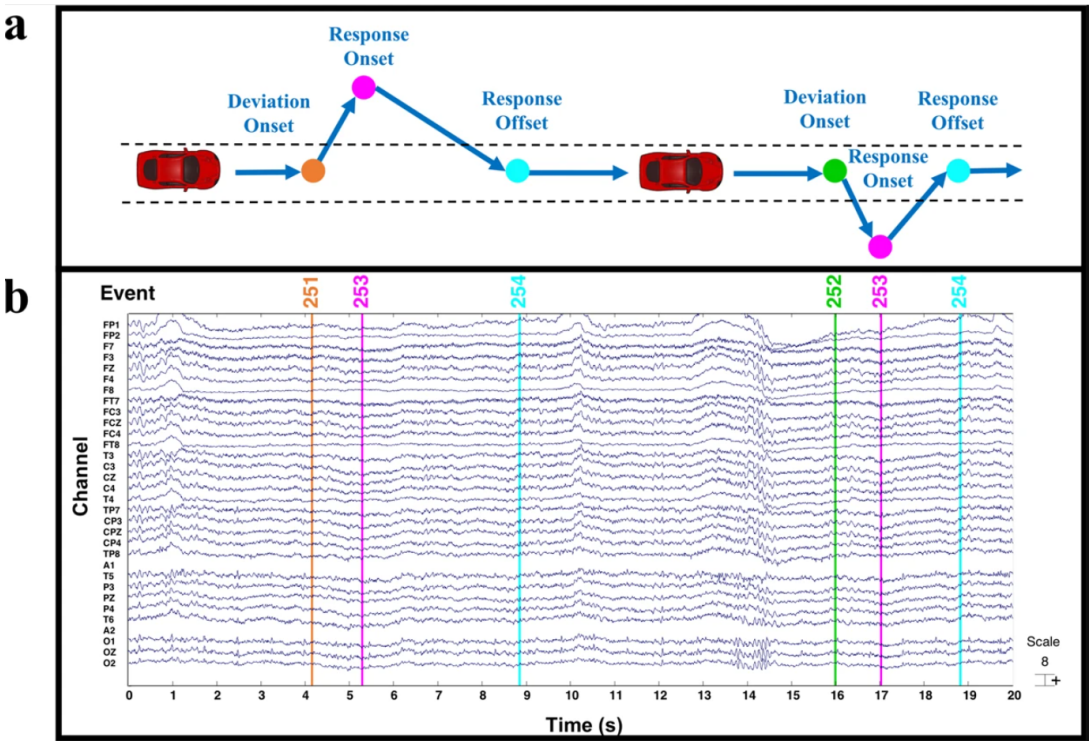


Figure 3.1: Experimental setup illustrating participants operating an immersive driving simulator with a motion platform. During a 90-minute night-time highway scenario, random lane departures were introduced to measure reaction times based on steering corrections [20].

was kept below 5 kΩ.

- **Behavioural Data:** Vehicle trajectory (lateral position) and event markers (deviation onset [left/right], response onset, response offset) were recorded by the simulation software and synchronized with the EEG data.

3.2.3 Data Availability and Relevance

This dataset [20] is publicly available via the figshare repository, provided in both raw and pre-processed formats (including artefact rejection and filtering), facilitating reproducibility and further research. Its primary relevance to this thesis lies in providing precisely timed EEG data locked to discrete events (lane deviations) with corresponding behavioural reaction times. This enables the investigation of pre-event neural activity as a predictor of subsequent driver performance, forming the basis for the reaction time prediction studies presented in Chapters 4 and 5.

3.3 Multimodal Driving Fatigue Dataset (Tobii)

The second dataset, utilized in Chapters 6 and 7, is a large-scale, internally collected multimodal dataset specifically designed to study driver drowsiness under controlled, yet realistic, conditions. Unlike the Cao et al. dataset, this dataset is not publicly available due to its proprietary nature and participant privacy considerations.

3.3.1 Participants and Fatigue Induction

Data were collected from 100 participants, with data from 79 individuals (age range 18-71 years; 30% female, 70% male) retained after quality control removed sessions with excessive noise or artefacts. A key aspect of this dataset is the controlled induction of fatigue. Each participant completed two driving simulation sessions:

1. **Alert State Session:** Conducted at 10 AM, following a normal night's sleep.

2. **Drowsy State Session:** Conducted at 3 AM, after the participant had remained awake for approximately 24 hours under supervision at the research facility.

This within-subject design, contrasting alert and sleep-deprived states, provides a strong basis for developing and evaluating drowsiness detection models. Subjective drowsiness was also assessed using the Karolinska Sleepiness Scale (KSS), confirming significantly higher drowsiness levels in the 3 AM sessions.

3.3.2 Data Acquisition and Modalities

Participants operated a high-fidelity driving simulator featuring a car seat, multiple screens, steering wheel, and pedals. The simulator environment was equipped with a suite of sensors to capture a wide range of physiological and behavioural data. While the full dataset includes EEG, EOG, EDA, EKG, SpO2, NIR imaging, RGB video, audio, and thermal IR imaging, the work presented in this thesis (Chapters 6 and 7) focuses specifically on:

- **EEG Data:** Recorded using Ag-AgCl electrodes, preprocessed to remove artefacts (e.g., using ICA), filtered into standard frequency bands (Delta, Theta, Alpha, Beta), and segmented into one-second non-overlapping windows.

- **Vision Data (RGB Video):** Captured by a standard camera pointing at the participant's face. Video frames were extracted and synchronized with the corresponding one-second EEG windows.

3.3.3 Data Handling and Relevance

Written informed consent was obtained from all participants, and strict data management protocols were followed to ensure confidentiality, including data anonymization and secure storage. The ground truth labels for the binary classification task ('Alert' vs. 'Drowsy') were determined directly from the experimental design: all data from the 10 AM sessions was labeled as **'Alert'**, and all data from the 3 AM sessions was labeled as **'Drowsy'**. To empirically validate this labeling strategy, subjective sleepiness ratings from the Karolinska Sleepiness Scale (KSS) [12], provided by an expert neurologist for each second of data, were analyzed. A statistical comparison confirmed that the mean KSS score at 3 AM was significantly higher than at 10 AM ($p < 0.00001$), providing strong validation for the ground truth labels. This robust labeling is critical for the relevance of this dataset, as it allows for the development and rigorous evaluation of advanced multimodal fusion models (Chapter 6) and the subsequent optimization of vision-based models for efficient edge deployment (Chapter 7).

3.3.4 Dataset Generation and Thesis Involvement

It is important to clarify the context of the Tobii dataset in relation to this thesis. The data collection and initial processing were undertaken by Tobii as a standalone, commercially focused project to create a large-scale, high-quality resource for internal research and development in driver monitoring. The generation of this dataset was not initially part of this PhD project but was a pre-existing, independent body of work.

My involvement began after the primary data collection phase. My role was to leverage this unique dataset to explore and develop the advanced machine learning and deep learning models that form the core of the multimodal and edge-deployment investigations (Chapters 6 and 7) in this thesis. The collaboration was synergistic; while analyzing the data, I provided feedback to Tobii regarding data quality, potential artefacts, and annotation consistency. This feedback loop contributed to the refinement of the final version of the dataset utilized for the research presented herein, ensuring its suitability for rigorous academic investigation while still aligning with its original commercial objectives.

3.4 Core Methodologies Explored

This thesis explores a range of signal processing and machine learning methodologies applied to the datasets described above, progressively increasing in complexity and tackling different aspects of driver state assessment. This section provides a high-level overview of these core techniques, which are detailed in subsequent chapters.

3.4.1 Pre-Event EEG for Reaction Time Prediction and Optimization

Chapter 4 focuses on the fundamental question of whether neural activity immediately preceding a critical event can predict subsequent behavioural response time and how such prediction can be optimized. Using the Cao et al. dataset [20], this work investigates:

- **Concept:** Predicting individual trial reaction times (RT) solely from EEG signals recorded in short pre-event windows.
- **Feature Extraction:** Power Spectral Density (PSD) estimation from pre-event EEG epochs in standard frequency bands.
- **Modeling Progression:** Starting with classical machine learning regression algorithms (Bayesian Ridge, ANNs), then introducing a 1D Convolutional Neural

Network (1D-CNN) for improved feature learning from spectral vectors.

- **Parameter Optimization:** Systematically investigating the impact of pre-stimulus window length, frequency band selection (individual vs. combined), and EEG channel subsets.
- **Evaluation Focus:** Subject-independent prediction accuracy (MAE, Pearson correlation) using Leave-One-Subject-Out (LOSO) cross-validation.

3.4.2 Vision-Based Analysis of EEG Spectral Images for Enhanced RT Prediction

Chapter 5 introduces an innovative approach to RT prediction by transforming EEG spectral features into 2D image representations (PSD Matrix Images and Scalp Topographies), still using the Cao et al. dataset [20]. This chapter explores:

- **Concept:** Leveraging powerful deep learning vision models by converting 1D EEG spectral data into 2D image formats.
- **Modeling Progression:** Initially evaluating a standard vision CNN (ResNet18 [97]) on these EEG-derived images, and then applying a more advanced Vision Transformer (ViT-B/16 [101]) to potentially capture global contextual patterns within these images more effectively.
- **Evaluation Focus:** Comparing the performance of these vision-based EEG pipelines (ResNet18 and ViT-B/16) against each other and, crucially, against the specialized 1D-CNN benchmark from Chapter 4, using subject-independent 5-fold cross-validation.

3.4.3 Multimodal Fusion for Drowsiness Classification using Transformers

Shifting focus from RT prediction to drowsiness classification, Chapter 6 utilizes the richer Tobii dataset (Section 3.3) to explore the fusion of EEG and visual (facial video) data:

- **Concept:** Developing models to classify driver state as 'Alert' or 'Drowsy' by combining synchronized EEG and video information.
- **Modeling Progression:** Establishing unimodal baselines (EEGNet [60] for EEG; ResNet18 and ViT-Base for vision), then exploring feature-level fusion (simple concatenation with Bayesian Ridge, and transformer-based feature fusion), and culminating in an end-to-end multimodal transformer architecture that processes raw EEG and video data jointly.
- **Evaluation Focus:** Subject-independent 5-fold cross-validation, using classification metrics (Accuracy, Precision, Recall, AUC-ROC) to assess the effectiveness of different fusion strategies compared to unimodal baselines.

3.4.4 Efficient Edge-Based Drowsiness Detection using Hybrid Transformers

The final empirical chapter, Chapter 7, addresses the practical challenge of deploying drowsiness detection systems on resource-constrained edge devices, focusing solely on the vision modality from the Tobii dataset:

- **Concept:** Designing an accurate yet computationally efficient deep learning model for real-time video-based drowsiness detection on mobile platforms.
- **Hybrid Architecture (MobileViT-LSTM):** Proposing a novel architecture combining an efficient vision transformer (MobileViT [114]) for per-frame spatial feature extraction and a Long Short-Term Memory network (LSTM [117]) for temporal aggregation over 5-second video windows.
- **Optimization for Edge:** Employing techniques like Automatic Mixed Precision (AMP) training [121] and model export to ONNX format [124].
- **Evaluation Focus:** Subject-independent 5-fold cross-validation using classification metrics, with a critical assessment of inference time on a representative edge

device to confirm real-time feasibility, compared against heavier ViT and standalone MobileViT baselines.

3.5 Evaluation Framework

A consistent and rigorous evaluation framework is employed across the empirical chapters (Chapters 4 through 7) to ensure the reliability and generalizability of the findings. The cornerstone of this framework is the adherence to **subject-independent validation**.

3.5.1 Subject-Independent Validation

Given that the ultimate goal is to develop systems applicable to unseen drivers without requiring extensive individual calibration, all models are evaluated using methods that strictly separate training and testing data at the participant level. This prevents the model from learning person-specific idiosyncrasies that do not generalize. Two primary strategies are used:

- **Leave-One-Subject-Out (LOSO) Cross-Validation:** Primarily used in Chapter 4. In each fold of the validation, data from one subject is held out entirely for testing, while the model is trained on data from all remaining subjects.

- **Subject-Independent k-Fold Cross-Validation:** Used in Chapters 5, 6, and 7 (typically with k=5). Participants are randomly partitioned into k folds. In each iteration, one fold of participants constitutes the test set, and the remaining k-1 folds of participants form the training set.

Performance metrics are typically averaged across all folds to provide a summary statistic, often accompanied by the standard deviation to indicate the variability in performance across different subject splits.

3.5.2 Performance Metrics

The choice of performance metrics is critical for a robust evaluation of the models and is directly informed by the standards and practices discussed in the state-of-the-art literature (Chapter 2). The metrics were selected to provide a comprehensive and task-appropriate assessment of model performance for both regression and classification problems.

Reaction Time Prediction (Regression - Chapters 4 and 5)

For the task of predicting the continuous RT value, the following two metrics were used, consistent with standard practice in regression and performance prediction literature:

- **Mean Absolute Error (MAE):** This metric measures the average absolute difference between the predicted and actual RTs, reported in seconds. MAE was chosen for its direct interpretability—it provides a clear and intuitive measure of the average prediction error in real-world units. It is also less sensitive to large outlier predictions compared to Mean Squared Error (MSE), which makes it a more robust metric for inherently variable behavioural data like RT.

- **Pearson Correlation Coefficient (r):** This metric measures the strength and direction of the linear relationship between the predicted and actual RTs. While MAE assesses absolute accuracy, the Pearson correlation assesses whether the model's predictions correctly track the trial-by-trial fluctuations in a driver's performance. As discussed in the literature review, achieving a significant positive correlation is a key indicator that a model has learned the underlying patterns of vigilance, even if its absolute predictions have a systematic offset. It is a crucial metric for validating the model's sensitivity to transient state changes.

Drowsiness Classification (Binary Classification - Chapters 6 and 7)

For the binary classification task of distinguishing 'Alert' from 'Drowsy' states, a suite of metrics was chosen to provide a holistic view of performance, which is essential for

safety-critical applications as highlighted in Chapter 2:

- **Accuracy:** The overall proportion of correct classifications. While a standard metric, it can be misleading if the classes are imbalanced.
- **Balanced Accuracy:** The average of recall for each class. This was selected as a more robust metric than standard accuracy, especially in cross-validation folds where slight class imbalances might occur, as it gives equal weight to the model's performance on both 'Alert' and 'Drowsy' states.
- **Precision and Recall (for the 'Drowsy' class):** These metrics are critical for safety applications. **Recall** (or Sensitivity) measures the model's ability to correctly identify truly drowsy instances, which is vital for minimizing false negatives (missed detections). **Precision** measures the proportion of drowsiness alerts that are correct, which is important for minimizing false positives and ensuring user trust in the system. The literature on driver monitoring systems frequently emphasizes the need to balance these two metrics.
- **Area Under the ROC Curve (AUC-ROC):** This is a comprehensive, threshold-independent measure of the model's ability to discriminate between the 'Alert' and 'Drowsy' classes. An AUC-ROC value provides a single scalar that summarizes the overall classification power of the model, making it an excellent metric for comparing different architectures, as is common practice in machine learning literature.

3.5.3 Baselines

Performance is consistently compared against relevant baselines to contextualize the results:

- Dummy Regressor/Classifier.
- Classical ML Models as baselines for more complex deep learning architectures.
- Unimodal Models as baselines for multimodal fusion.

- Standard (heavier) Architectures as baselines for proposed efficient models.

This comprehensive evaluation framework ensures that the contributions of the different methodologies are assessed rigorously in terms of both accuracy and generalizability, and in the final stages, practical deployment feasibility.

3.6 Summary

This chapter has laid the groundwork for the empirical investigations within this thesis by detailing the two key datasets employed: the Cao et al. [20] public EEG dataset for reaction time studies and the internally collected Tobii multimodal dataset for drowsiness classification. It has also provided a roadmap of the core methodologies explored, tracing a path from classical machine learning on spectral EEG features, through advanced CNN and transformer-based approaches for both unimodal and multimodal data, and culminating in the development of efficient hybrid models tailored for edge deployment. Finally, the chapter outlined the rigorous, subject-independent evaluation framework and associated metrics that underpin the assessment of model performance and generalizability throughout the subsequent chapters. With this context established, the following chapters will now delve into the specific implementations, results, and discussions of each methodological stage.

Chapter 4

Electroencephalography-Based Prediction of Driver Reaction Time using Pre-Stimulus Neural Activity

4.1 Introduction

Ensuring driver safety is an important global challenge, with driver states such as fatigue, drowsiness, and inattention being major contributors to road accidents [2, 3]. A direct and often critical consequence of impaired driver state is an increase in reaction time (RT) when responding to unexpected or hazardous road events [5]. Slowed reactions significantly reduce the time available for corrective maneuvers, thereby increasing crash risk. While various methods exist to assess driver state, many are subjective (e.g., questionnaires [7]) or require active participation (e.g., vigilance tests [8]), making them unsuitable for continuous, non-intrusive monitoring during driving.

Physiological signals offer a promising avenue for objective, real-time assessment. Among these, Electroencephalography (EEG) stands out as it provides a direct measure of cortical brain activity, reflecting the dynamic changes in cognitive states like alertness, attention, and cognitive load that fundamentally underpin driving performance [34, 35, 36]. The richness of the EEG signal holds the potential not just to classify a driver's

current state but perhaps even to anticipate their near-future behaviour.

This chapter delves into this anticipatory potential, addressing the core challenge: can I predict an individual's reaction time to an imminent event using only the neural activity captured via EEG in the moments immediately preceding that event? Such a capability could form the basis for proactive Advanced Driver Assistance Systems (ADAS) that could adapt or intervene before a potentially dangerous slow response occurs.

Specifically, this chapter aims to systematically investigate the feasibility and subsequently optimize the parameters for predicting driver reaction time to simulated lane-departure events based solely on pre-stimulus EEG spectral features. I focus on a subject-independent framework, ensuring that my findings are generalizable to unseen individuals, a crucial requirement for practical applications. This investigation is guided by two primary research questions:

1. **RQ1: Can pre-stimulus EEG spectral features, particularly from alpha and theta bands, reliably predict driver reaction time to unexpected lane deviation events in a subject-independent framework?**

   I investigate whether power spectral density features extracted from short, pre-event EEG windows (-2s to 0s) can effectively differentiate fast from slow responses on a trial-by-trial basis and estimate average response tendencies, using standard machine learning pipelines evaluated rigorously across participants.

2. **RQ2: Can the predictive performance be enhanced by systematically optimizing EEG input parameters (pre-stimulus window length, frequency band selection, channel subsets) and employing more advanced feature learning models (1D-CNN)?**

   I explore the impact of varying key input parameters and compare the performance of classical machine learning models against a 1D Convolutional Neural Network designed to automatically learn relevant patterns from the spectral features, seeking to maximize predictive robustness.

To address these questions, I utilize the publicly available multi-channel EEG driving

dataset collected by Cao et al. [20] (detailed in Chapter 3), which provides synchronized EEG and behavioural RT data during a sustained attention task. My analysis focuses on a subset of these participants, excluding data compromised by significant artefacts or anomalous behaviour to ensure the robustness of my findings. This chapter progresses from establishing the basic feasibility of pre-event RT prediction using classical machine learning models (addressing RQ1) to systematically exploring input parameters and leveraging a 1D-CNN architecture for improved performance (addressing RQ2), culminating in an interpretable analysis of the underlying neural patterns driving predictability.

4.2 Methods

4.2.1 Dataset and Experimental Paradigm

The analysis presented in this chapter is based on the publicly available dataset described by Cao et al. [20], which was introduced in detail in Section 3.2. Briefly, the dataset comprises synchronized 32-channel EEG recordings and behavioural data from participants engaged in a 90-minute simulated driving task. The core task involved maintaining lane position on a monotonous highway, during which random lane deviation events were introduced. Participants were required to make corrective steering actions. The key event markers relevant to this study are the 'Deviation Onset' (start of the drift, codes 251 for left, 252 for right) and 'Response Onset' (start of corrective steering, code 253). The Reaction Time (RT) for each trial is defined as the temporal difference between the Deviation Onset and the corresponding Response Onset.

4.2.2 Participant Subset Selection and Justification

The original dataset published by Cao et al. [20] included data from 27 participants across 62 sessions. For the analyses conducted in this chapter, I utilized data from 24 of these participants. Three participants were excluded following careful inspection of their EEG data and behavioural performance. This exclusion was based on two primary criteria:

1. **Excessive EEG Artefacts:** Several recording sessions from the excluded participants exhibited substantial contamination from non-neural sources, such as excessive muscle activity (electromyography, EMG) or significant movement artefacts. Such artefacts can severely distort the underlying EEG signal, compromising the integrity of spectral feature extraction and potentially leading to unreliable model training [126]. Standard preprocessing techniques may not fully mitigate severe or pervasive artefacts [127].

2. **Anomalous Reaction Time Distributions:** The behavioural data for these participants displayed highly atypical reaction time patterns in some sessions. This included disproportionately long RTs (significantly exceeding the typical range observed in the majority of participants) or extreme variability in RTs across trials. Such patterns might indicate periods of task disengagement, misunderstanding of instructions, or potentially microsleeps or extreme fatigue states that deviate significantly from the typical alert-to-drowsy continuum the models aim to capture. Including these anomalous data points could unduly bias the regression models, leading them to focus on predicting extreme outliers rather than the more subtle, continuous variations in RT associated with typical fluctuations in vigilance and attention.

The exclusion of these participants aligns with standard practices in EEG research, aiming to enhance the signal-to-noise ratio and ensure that the machine learning models are trained on data representative of the cognitive processes under investigation (i.e., variations in attention and preparedness influencing typical reaction speeds). This focus on a cleaner, more representative subset of the data allows for a more robust assessment of the relationship between pre-stimulus EEG and typical driver reaction time variability. All subsequent analyses reported in this chapter are based on the data from the remaining 24 participants.

4.2.3 EEG Data Preprocessing and Epoching

The raw EEG data (‘.set‘ files from the Cao et al. dataset [20]) for the selected 24 participants were processed using the MNE-Python library [56]. Initial preprocessing steps involved removing non-EEG channels (e.g., the ’vehicle_position’ channel) and applying a standard 10-20 montage definition to ensure spatial consistency across recordings.

A crucial step was filtering the continuous EEG data to remove noise and baseline drift while preserving physiologically relevant frequencies. A Finite Impulse Response (FIR) band-pass filter was applied between 0.5 Hz and 30 Hz. This frequency range encompasses the delta, theta, alpha, and lower beta bands, which are commonly associated with cognitive states like alertness, attention, and drowsiness [14]. The filtering was performed using a zero-phase Hamming window FIR (Finite Impulse Response) filter.

Following filtering, the continuous data was segmented into epochs time-locked to the ’Deviation Onset’ events (codes 251 or 252). To investigate the predictive power of pre-stimulus activity (addressing RQ1 and RQ2), epochs were extracted from the time window immediately preceding the deviation onset. While the primary analysis focused on a 2-second pre-stimulus window (‘tmin=-2.0‘, ‘tmax=0.0‘), the systematic investigation for RQ2 also evaluated window lengths of 1, 3, 4, and 5 seconds (‘tmin‘ ranging from -1.0 to -5.0, ‘tmax=0.0‘). No baseline correction was applied to these pre-stimulus epochs, as the interest lies in the absolute signal characteristics within this specific window.

The target variable for prediction, Reaction Time (RT), was calculated for each valid trial as the time difference between the ’Response Onset’ marker (code 253) and the preceding ’Deviation Onset’ marker (code 251 or 252). This difference, measured in samples, was divided by the sampling rate (500 Hz) to obtain RT in seconds. Trials with RTs falling outside a physiologically plausible range were excluded from further analysis. Based on typical human reaction times in driving contexts and inspection of the data distribution, trials with RTs shorter than 0.2 seconds or longer than 5.0 seconds were removed. This step helps to eliminate potential outliers caused by accidental responses, lapses in attention, or measurement errors, ensuring the models focus on predicting typical variations in response speed. The RT values corresponding to the selected valid epochs

were stored as the target variable ‘y_data‘.

4.2.4 Feature Extraction: Power Spectral Density (PSD)

To quantify the frequency-specific neural activity within the pre-stimulus epochs, Power Spectral Density (PSD) was estimated using Welch's method [57]. This is a standard technique in EEG analysis that provides a robust estimate of the power distribution across different frequencies by averaging modified periodograms computed on overlapping segments of the signal [128].

PSD was computed for each channel within each pre-stimulus epoch using the ‘mne.time_frequency.ps function. Key parameters for Welch's method were set as ‘n_fft=1000‘ (length of the Fast Fourier Transform window, determining frequency resolution) and ‘n_overlap=500‘ (number of samples overlapping between consecutive windows). Power values were then averaged within specific canonical frequency bands, defined as:

- Delta (δ): 0.5 – 4 Hz
- Theta (θ): 4 – 8 Hz
- Alpha (α): 8 – 12 Hz
- Beta (β): 14 – 20 Hz

The Gamma band (typically >30 Hz) was excluded due to the 30 Hz low-pass filter applied during preprocessing and its higher susceptibility to muscle artefacts.

As part of the investigation for RQ2, in addition to these individual bands, PSD features were also computed for two combined arithmetic measures previously explored in cognitive research [14, 67, 5]:

- Theta+Beta / Alpha: Sum of theta and beta power, divided by alpha power.
- Theta / Beta: Ratio of theta power to beta power.

For each epoch and each frequency band (or combination), the PSD values computed across the 32 EEG channels were concatenated (flattened) into a single feature vector.

This resulted in a feature matrix where each row corresponds to a pre-stimulus epoch (trial) and each column represents the PSD value for a specific channel within the chosen frequency band. These spectral features formed the input for the machine learning models described below.

4.2.5 Exploration of Input Parameters

To systematically address RQ2 and optimize the prediction pipeline, I investigated the influence of several key input parameters on model performance:

1. **Pre-stimulus Window Length:** The duration of the EEG segment immediately preceding the deviation onset was varied. I tested windows of 1 second (-1.0s to 0.0s), 2 seconds (-2.0s to 0.0s), 3 seconds (-3.0s to 0.0s), 4 seconds (-4.0s to 0.0s), and 5 seconds (-5.0s to 0.0s). The goal was to identify the optimal duration that balances capturing sufficient predictive neural activity against introducing excessive noise or irrelevant information from further back in time [49].

2. **Frequency Band Selection:** I compared the predictive performance using PSD features derived from each individual band (Delta, Theta, Alpha, Beta) against the performance using the combined arithmetic measures (Theta+Beta/Alpha, Theta/-Beta). This aimed to determine whether specific bands hold unique predictive value or if combined indices offer advantages [5, 73].

3. **Channel Subsets:** Recognizing that practical applications might benefit from reduced sensor configurations, I evaluated model performance using PSD features extracted from spatially distinct subsets of EEG channels, compared to using the full 32-channel set. Based on the standard 10-20 system locations provided in the dataset [20], channels were grouped into the following regional subsets:

    - **Frontal:** Fp1, Fp2, F7, F3, Fz, F4, F8
    - **Central:** C3, Cz, C4
    - **Temporal:** T3, T4, T5, T6

- **Parietal:** P3, Pz, P4
- **Occipital:** O1, Oz, O2

The analysis compared these regional subsets against the full 32-channel configuration. This investigation aimed to identify if specific brain regions were particularly crucial for RT prediction or if redundancy exists, potentially allowing for simpler EEG setups [86, 78, 82].

The results of these explorations informed the selection of optimal parameters for the final model comparisons.

4.2.6 Machine Learning Models

To predict RT from the extracted PSD features, I employed and compared several machine learning models:

1. **Baseline Model ('DummyRegressor'):** A simple baseline was established using Scikit-learn's 'DummyRegressor' configured with the 'mean' strategy. This model simply predicts the average RT observed in the training dataset for every trial in the test set. It serves as a benchmark to determine if the more complex models learn any meaningful predictive patterns beyond the overall average response time.
2. **Classical Models:**
   - **Bayesian Ridge Regression ('BayesianRidge'):** A linear regression model that incorporates Bayesian inference with Gaussian priors on the model weights [109]. This provides regularization, helping to prevent overfitting, particularly when the number of features (channels x frequency bins) might be relatively high compared to the number of trials for some subjects. It was implemented using Scikit-learn with default hyperparameters.
   - **Artificial Neural Network (ANN / 'MLPRegressor'):** A shallow Multi-Layer Perceptron (MLP) was used as implemented in Scikit-learn's 'MLPRegressor'. The architecture consisted of three hidden layers, each with 25 neu-

rons. The hyperbolic tangent ('tanh') activation function was used for hidden layers. Regularization was applied using an L2 penalty (alpha = 0.01). The model was trained for a maximum of 500 iterations using the 'adam' optimizer with a learning rate of 0.001 and Nesterov's momentum enabled. Input features (PSDs) were standardized (zero mean, unit variance) using 'StandardScaler' fit only on the training data within each cross-validation fold.

3. **Advanced Model (1D-CNN):** To explore the potential of deep learning for automatically extracting relevant patterns from the spectral feature vectors, a one-dimensional Convolutional Neural Network (1D-CNN) was designed and implemented, using the PyTorch framework. This architecture is referred to as **specialised** because, unlike a general-purpose fully connected network (ANN), its core components are specifically designed for processing 1D sequential or spatio-temporal data like the ordered vector of EEG channel features. The 1D convolutional layer possesses strong inductive biases—namely **locality** (assuming adjacent channels in the vector are related) and **parameter sharing** (applying the same pattern detector across the entire vector)—which make it highly effective at learning localized spatial patterns that are characteristic of neurophysiological signals. The architecture consisted of:

   - An initial 1D convolutional layer responsible for learning local patterns across the spectral features. Hyperparameter tuning identified 64 filters with a kernel size of 3 as optimal.
   - A Rectified Linear Unit (ReLU) activation function following the convolutional layer. ReLU was found to perform better than 'tanh', potentially due to faster convergence and mitigation of vanishing gradients.
   - A flattening layer to convert the output of the convolutional layer into a 1D vector.
   - A final dense (fully connected) layer with a single output neuron (linear activation) to produce the continuous RT prediction.

The 1D-CNN was trained using the Mean Absolute Error (MAE) as the loss function, which directly optimizes the primary evaluation metric. The 'adam' optimizer was employed. This architecture aims to capture potentially complex, non-linear relationships and interactions within the PSD features that might be missed by the linear Bayesian Ridge or the shallow ANN [58, 129].

All models were trained and evaluated within the subject-independent cross-validation framework described below.

4.2.7 Evaluation Strategy

The core evaluation strategy employed throughout this chapter was designed to rigorously assess the models' ability to generalize to unseen individuals, reflecting a crucial requirement for practical applications. I used **Leave-One-Subject-Out (LOSO) cross-validation** across the 24 selected participants.

In this procedure:

1. The dataset was iteratively split 24 times.

2. In each iteration (fold), the data from a single participant was completely held out as the test set.

3. The model was trained using the combined data from the remaining 23 participants.

4. Feature scaling (using 'StandardScaler') was performed separately within each fold, fitting the scaler only on the training data (23 subjects) and then applying it to both the training and the held-out test data.

5. The trained model was used to predict RTs for all trials belonging to the held-out test subject.

6. Performance metrics were calculated by comparing the predictions against the actual RTs for the test subject.

This process was repeated for all 24 subjects, resulting in 24 sets of performance metrics. The final reported performance measures represents the mean and standard deviation of these metrics across the 24 folds, providing an estimate of the expected performance on a new, unseen individual, along with the inter-subject variability.

The primary performance metrics used were:

- **Mean Absolute Error (MAE):** Calculated as the average absolute difference between the predicted RT (RT_{pred}) and the actual RT (RT_{true}) for the test subject in each fold:
$$MAE = \frac{1}{N_{test}} \sum_{i=1}^{N_{test}} |RT_{pred,i} - RT_{true,i}|$$
where N_{test} is the number of trials for the test subject. Lower MAE indicates better predictive accuracy. MAE is reported in seconds.

- **Pearson Correlation Coefficient (r):** Calculated between the vector of predicted RTs and the vector of actual RTs for the test subject in each fold. It measures the strength and direction of the linear relationship between predictions and ground truth. Values range from -1 to 1. A positive correlation indicates that the model tends to predict higher RTs for trials where the actual RT was higher, capturing the trial-by-trial fluctuations. The statistical significance of the correlation (p-value) was also assessed (e.g., $p < 0.05$ or $p < 0.01$ indicating significance).

4.2.8 Interpretability: Common Spatial Patterns (CSP)

To gain insight into the neurophysiological underpinnings of the RT predictions and to visually inspect the scalp projections contributing most to distinguishing fast from slow responses, the Common Spatial Patterns (CSP) algorithm [87] was employed. CSP is a spatial filtering technique widely used in Brain-Computer Interfaces (BCIs) to find linear combinations of EEG channels (spatial filters) that maximize the variance for one condition (e.g., slow RTs) while minimizing it for another (e.g., fast RTs).

Since CSP requires binary class labels, the continuous RT values were first categorized into 'high RT' and 'low RT' groups for each subject. This was achieved by computing the

Z-score of the RTs within each subject's data. Trials with a positive Z-score (RT above the subject's mean RT) were labeled as 'high RT', and trials with a negative Z-score (RT below the subject's mean RT) were labeled as 'low RT'.

The CSP algorithm was then applied to the pre-stimulus EEG epoch data (typically filtered in a relevant band like alpha or theta) using these derived binary labels. The algorithm yields a set of spatial filters (weight vectors across channels). The patterns associated with these filters (obtained by applying the inverse of the filter matrix) represent the scalp topographies that show the strongest differences between the high and low RT conditions.

These resulting spatial patterns were visualized as scalp topographies, typically showing the two most discriminative patterns (one maximizing variance for high RT, the other for low RT). Red areas in these topographies indicate electrode locations with high positive weights in the pattern, suggesting these regions contribute significantly to the variance differences captured by the filter. This visualization helps to interpret the model's predictions in a neurophysiologically meaningful way, confirming whether the predictive features originate from plausible brain sources associated with attention, motor preparation, or vigilance, rather than potential artefacts [88].

4.3 Results

This section presents the empirical findings from applying the methodologies described above to the pre-stimulus EEG data of the 24 selected participants. I first address the feasibility of predicting reaction time using classical models (RQ1) and then detail the results of parameter optimization and the performance enhancement achieved with the 1D-CNN model (RQ2).

4.3.1 Feasibility of Pre-Event Reaction Time Prediction (RQ1)

The initial analyses focused on establishing whether pre-event EEG spectral features contain sufficient information to predict subsequent reaction times in a subject-independent

manner, using the classical ANN and Bayesian Ridge models. For these initial feasibility results, I anticipate the findings from the parameter optimization (Section 4.3.2) and primarily use the 2-second pre-stimulus window and features from the Alpha (8-12 Hz) and Theta (4-8 Hz) bands, as these were found to be most effective.

Trial-Level Reaction Time Prediction Accuracy

The primary goal was to assess if the models could predict the RT for individual upcoming lane deviation events based solely on the preceding 2 seconds of EEG. Table 4.1 presents the Mean Absolute Error (MAE), averaged across the 24 subjects using LOSO cross-validation, comparing the performance of Bayesian Ridge and ANN models against the Dummy Regressor baseline. Lower MAE values indicate more accurate predictions.

Table 4.1: Aggregate Mean Absolute Errors (MAE in seconds, mean ± std dev) across 24 subjects for trial-level RT prediction using subject-independent classical models (2-second pre-stimulus window). Lower values indicate better accuracy.

Frequency Band	Bayesian Ridge	ANN (MLP Regressor)	Dummy Regressor
Alpha (8-12 Hz)	0.53 ± 0.25	**0.51 ± 0.23**	0.58 ± 0.27
Theta (4-8 Hz)	0.55 ± 0.32	0.54 ± 0.29	0.58 ± 0.27
Beta (14-20 Hz)	0.58 ± 0.26	0.59 ± 0.26	0.58 ± 0.27
Delta (0.5-4 Hz)	0.57 ± 0.27	0.54 ± 0.26	0.58 ± 0.27

As shown in Table 4.1, both the Bayesian Ridge and ANN models, when utilizing features from the alpha, theta, or delta bands, achieved lower average MAEs than the Dummy Regressor baseline (average MAE $\approx$ 0.58s). This demonstrates that the models successfully learned predictive patterns from the pre-event EEG beyond simply predicting the overall mean RT. The best performance was achieved by the ANN model using alpha band features, yielding an average MAE of 0.51s, approximately a 12% reduction compared to the baseline. Theta and delta bands also provided predictive information, particularly with the ANN. The beta band features, however, offered little to no predictive advantage over the baseline. The standard deviations (0.23s to 0.32s) indicate considerable inter-subject variability in prediction accuracy, a common characteristic in subject-independent EEG analysis.

Figure 4.1 provides a visual comparison of the MAE distributions across subjects for the best model configuration per band versus the Dummy baseline.

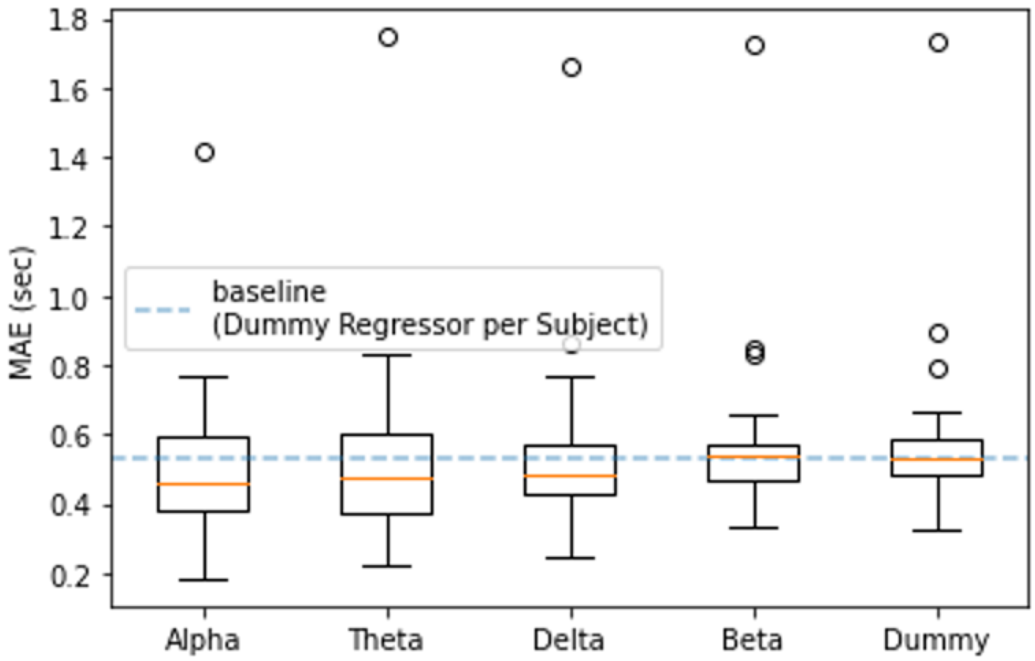


Figure 4.1: Boxplot comparing MAE distributions across 24 subjects for trial-level RT prediction using classical models. Each box represents the results using the best model (ANN or Bayesian Ridge) per EEG frequency band (Alpha, Theta, Delta, Beta) compared against the Dummy Regressor baseline. The central line is the median, the box spans the interquartile range (IQR), and whiskers typically extend to 1.5x IQR.

The boxplot visually confirms the findings from the table. The median MAE for models trained on alpha, theta, and delta features is clearly lower than the Dummy Regressor's median. The distributions for alpha and theta models appear shifted downwards compared to the baseline, indicating better performance for the majority of subjects. The spread of the boxes and whiskers underscores the substantial variability between individuals.

Per-Subject Performance and Correlation

To assess the consistency of prediction success at the individual level, Table 4.2 details the performance for each of the 24 subjects when their data served as the held-out test set in the LOSO validation. It reports the MAE achieved by the ANN model (generally the better classical model) using Alpha and Theta features, along with the Pearson

correlation coefficient ('corr') between the model's predictions and the subject's actual RTs. Significant correlations indicate that the model captured meaningful trial-by-trial RT fluctuations.

The per-subject results confirm the high degree of individual variability. MAE values range significantly (e.g., from 0.18s for Sub 27 with alpha features to 1.40s for Sub 8). However, a crucial finding is that the Pearson correlations between predicted and actual RTs are statistically significant ($p<0.05$ or $p<0.01$) for the vast majority of subjects when using either alpha or theta band features. This indicates that, even when the absolute prediction error (MAE) is relatively high for some individuals compared to their own average RT (compare MAE column to RT(avg)), the model is successfully capturing the trial-to-trial variance in their reaction speed based on the pre-event EEG. For example, Subject 3 exhibits a high MAE ($\approx$0.8s) but also a strong, highly significant correlation ($r=0.40^{**}$ for alpha), meaning the model's predictions tracked the ups and downs of their actual RTs, even if systematically offset. This consistent ability to predict within-subject RT fluctuations for unseen individuals provides strong support for the feasibility of subject-independent pre-event RT prediction (RQ1).

Predicting Average Reaction Time Tendency

Beyond predicting individual trial RTs, I investigated whether the subject-independent models could discern a participant's general response tendency (i.e., whether they are typically a fast or slow responder). Figure 4.2 plots the average RT predicted by the ANN alpha model for each subject (when they were the test subject) against their actual average RT calculated across all their valid trials.

The strong positive linear correlation ($r = 0.71$, $p < 0.0001$) observed in Figure 4.2 is a significant finding. It demonstrates that the subject-independent model, despite having never been trained on the specific test individual's data, can reliably estimate their characteristic average reaction speed relative to the other participants. Individuals who were generally slower responders (higher actual average RT) consistently received higher average predicted RTs from the model, and vice versa. This indicates that stable,

Table 4.2: Subject-independent prediction results per subject using the ANN (MLP Regressor) model with Alpha and Theta features (2-second window). N: number of valid trials, corr: Pearson correlation ($p<0.05$*, $p<0.01$**), MAE (s), RT (avg): subject's average actual RT (s), Dummy MAE: baseline MAE for the subject.

Sub	N	Alpha Features		Theta Features		RT	Dummy
		corr	MAE	corr	MAE	(avg)	MAE
1	780	0.04	0.45	0.05	0.49	1.24	0.38
2	673	0.12**	0.31	0.10**	0.36	0.78	0.49
3	356	0.40**	0.77	0.29**	0.83	1.53	0.89
4	1356	0.33**	0.64	0.18**	0.68	1.15	0.55
5	355	0.34**	0.48	0.31**	0.43	1.13	0.43
6	617	0.06	0.48	0.07	0.53	1.03	0.51
7	414	0.28**	0.40	0.22**	0.37	0.78	0.59
8	499	0.25**	1.40	0.17**	1.74	2.57	1.73
9	737	0.12**	0.47	0.18**	0.55	0.69	0.60
10	727	0.07*	0.36	0.04	0.22	0.58	0.60
11	1412	0.09**	0.67	0.08**	0.70	1.43	0.67
12	434	0.31**	0.30	0.31**	0.32	1.07	0.32
13	1173	0.11**	0.42	0.12**	0.52	1.40	0.42
14	983	0.20**	0.42	0.27**	0.45	1.00	0.49
15	2031	0.07	0.58	0.37**	0.54	1.30	0.53
16	748	0.29**	0.76	0.35**	0.80	1.86	0.79
17	2234	0.13**	0.58	0.04*	0.69	1.09	0.56
18	330	0.18*	0.36	0.38**	0.35	0.89	0.55
19	1007	0.25**	0.38	0.27**	0.37	0.88	0.49
20	669	0.36**	0.41	0.38**	0.40	1.09	0.40
21	205	0.54**	0.28	0.58**	0.26	0.91	0.50
22	1094	0.16**	0.55	0.06	0.57	1.28	0.56
23	164	0.06	0.63	0.40**	0.45	1.41	0.53
24	637	0.04	0.18	0.02*	0.29	0.71	0.58

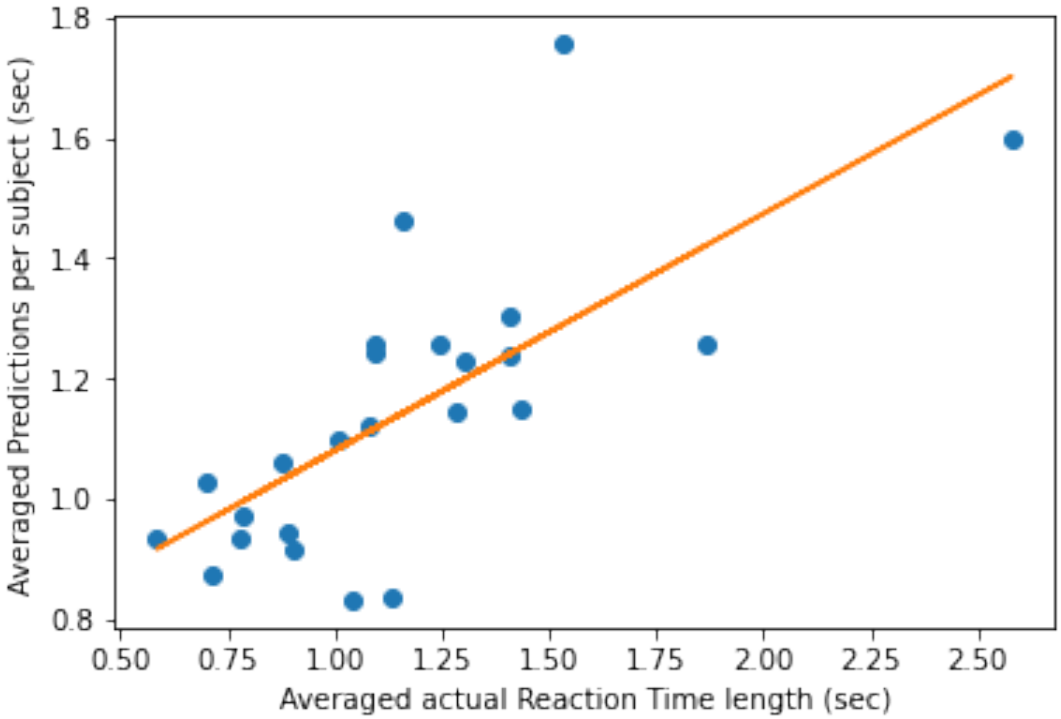


Figure 4.2: Relationship between average actual reaction time (Ground Truth RT) and average predicted reaction time per subject, using the ANN alpha model (N=24 subjects). Each point represents one subject. The strong positive linear correlation (Pearson's r=0.71, p<0.0001) indicates the subject-independent model effectively differentiates between generally fast and slow responders.

inter-individually consistent EEG patterns related to overall response speed are present in the pre-event window and can be learned effectively by the model. This capability complements the trial-level prediction, offering a way to potentially classify drivers based on their general responsiveness profile derived purely from pre-event EEG.

Relationship between Prediction Error and Actual RT

To explore factors influencing prediction accuracy, I examined the relationship between the model's prediction error (MAE) for a given subject and that subject's average actual RT. Figure 4.3 plots the per-subject MAE from the ANN alpha model against their average RT.

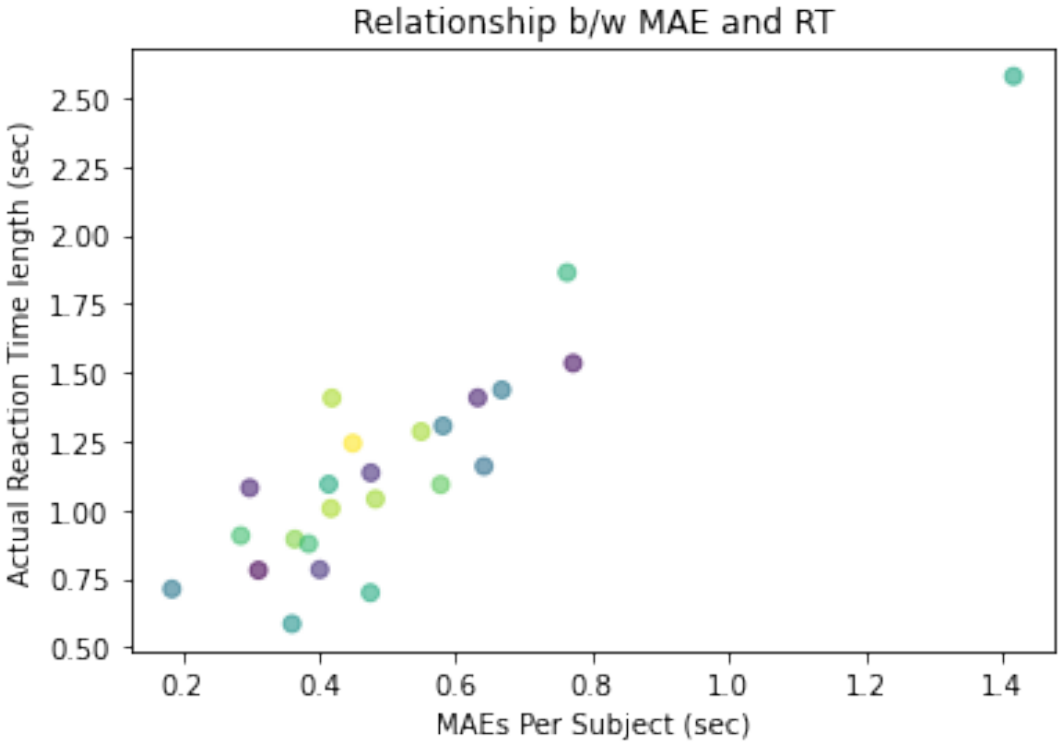


Figure 4.3: Scatter plot showing the relationship between the per-subject prediction Mean Absolute Error (MAE) from the ANN alpha model and the subject's average actual Reaction Time (RT). Each point represents one subject (N=24). A positive trend suggests predictions are generally less accurate for slower responders.

Figure 4.3 reveals a discernible positive trend: subjects with generally longer average reaction times tend to have higher prediction errors (MAE). This suggests that predicting the precise RT from pre-event EEG is more challenging for individuals who are inherently

slower or perhaps exhibit greater variability in their responses. Faster responders might present clearer or more stable pre-event neural signatures associated with their RTs, making them more predictable by the current models and features. This finding has implications for model development, potentially indicating the need for different strategies or features to accurately predict RTs across the full spectrum of individual response speeds.

Collectively, these results strongly support an affirmative answer to RQ1: pre-stimulus EEG features, particularly from the alpha and theta bands, do contain information predictive of subsequent reaction time, enabling subject-independent models to outperform baseline predictions both at the trial level (as shown by significant correlations for most subjects) and in estimating overall response tendencies.

4.3.2 Optimization of Input Parameters (RQ2)

Having established the feasibility of pre-event RT prediction, I systematically investigated how different input parameters affect performance, aiming to optimize the pipeline (addressing RQ2). These analyses primarily used the ANN model, which generally outperformed Bayesian Ridge.

Effect of Pre-stimulus Window Length

I evaluated the impact of varying the pre-stimulus window length from 1 second to 5 seconds on prediction accuracy (MAE) and correlation (PCC). Table 4.3 summarizes the MAE results, and Table 4.4 presents the corresponding Pearson correlations, averaged across the 24 subjects.

The results indicate that the choice of window length influences performance. Based on MAE (Table 4.3), the **2-second pre-stimulus window** consistently yields the best or near-best performance, particularly for the most predictive Alpha and Theta bands. Shorter windows (1s) might not capture enough evolving neural state information, while longer windows (3s-5s) seem to introduce noise or irrelevant past activity, sometimes degrading performance below the 2s level (e.g., Alpha MAE increases significantly at

Table 4.3: Comparison of Mean Absolute Errors (MAE, mean ± std dev) across frequency bands for the ANN model using different pre-stimulus window lengths (-1s to -5s). Improvement percentages relative to the Dummy Regressor (MAE=0.58) are in parentheses.

Band	**1-sec**	**2-sec**	**3-sec**	**4-sec**	**5-sec**
Alpha	0.55±0.35 (5%)	**0.51±0.23 (12%)**	0.67±0.36 (-16%)	0.54±0.37 (7%)	0.57±0.28 (2%)
Theta	0.55±0.35 (5%)	**0.54±0.33 (7%)**	**0.54±0.32 (7%)**	**0.54±0.26 (7%)**	0.56±0.37 (3%)
Beta	0.61±0.33 (-5%)	0.59±0.30 (-2%)	0.61±0.32 (-5%)	0.60±0.27 (-3%)	0.56±0.31 (3%)
Delta	0.61±0.28 (-5%)	**0.54±0.28 (7%)**	0.55±0.33 (5%)	**0.54±0.45 (7%)**	0.65±0.30 (-12%)

Table 4.4: Comparison of Pearson Correlation Coefficients (PCC, mean ± std dev) across frequency bands for the ANN model using different pre-stimulus window lengths.

Band	**1-sec**	**2-sec**	**3-sec**	**4-sec**	**5-sec**
Alpha	0.18±0.14	**0.21±0.16**	0.20±0.16	0.18±0.17	0.19±0.16
Theta	0.20±0.12	0.24±0.13	**0.26±0.13**	0.25±0.12	0.23±0.13
Beta	0.07±0.10	0.06±0.12	0.08±0.13	0.11±0.11	0.10±0.13
Delta	0.17±0.11	0.21±0.13	**0.23±0.13**	**0.23±0.14**	0.21±0.13

3s and 5s). The correlation results (Table 4.4) show a similar trend, with correlations generally peaking around the 2s to 4s window lengths for the Theta and Delta bands, and at 2s for the Alpha band. Considering both MAE and correlation, the 2-second window appears to offer an optimal balance, capturing critical predictive information immediately preceding the event without being overly contaminated by noise or less relevant prior activity. Consequently, the 2-second window was adopted for subsequent analyses unless otherwise specified.

Effect of Frequency Bands

Using the optimal 2-second window, I compared the predictive power of individual frequency bands against the combined arithmetic measures (Theta+Beta/Alpha and Theta/-Beta). Table 4.5 summarizes the MAE results for the ANN model.

Figure 4.4 provides a visual comparison of the MAE distributions.

The results clearly demonstrate that the individual **Alpha and Theta bands contain the most predictive information** for RT in this task. They yield the lowest average MAEs and show the largest improvement over the Dummy baseline. The combined band measures (Theta+Beta/Alpha, Theta/Beta), while slightly better than the

Table 4.5: Aggregate MAEs (mean ± std dev) for the ANN model using individual vs. combined frequency band features (2-second window). Improvement percentages relative to Dummy (MAE=0.58) in parentheses.

Frequency Band / Combination	**ANN MAE**	**Dummy MAE**
Alpha	**0.51 ± 0.23 (12%)**	0.58 ± 0.27
Theta	**0.54 ± 0.29 (7%)**	0.58 ± 0.27
Beta	0.59 ± 0.26 (-2%)	0.58 ± 0.27
Delta	0.54 ± 0.26 (7%)	0.58 ± 0.27
Theta+Beta/Alpha	0.56 ± 0.24 (3%)	0.58 ± 0.27
Theta/Beta	0.57 ± 0.24 (2%)	0.58 ± 0.27

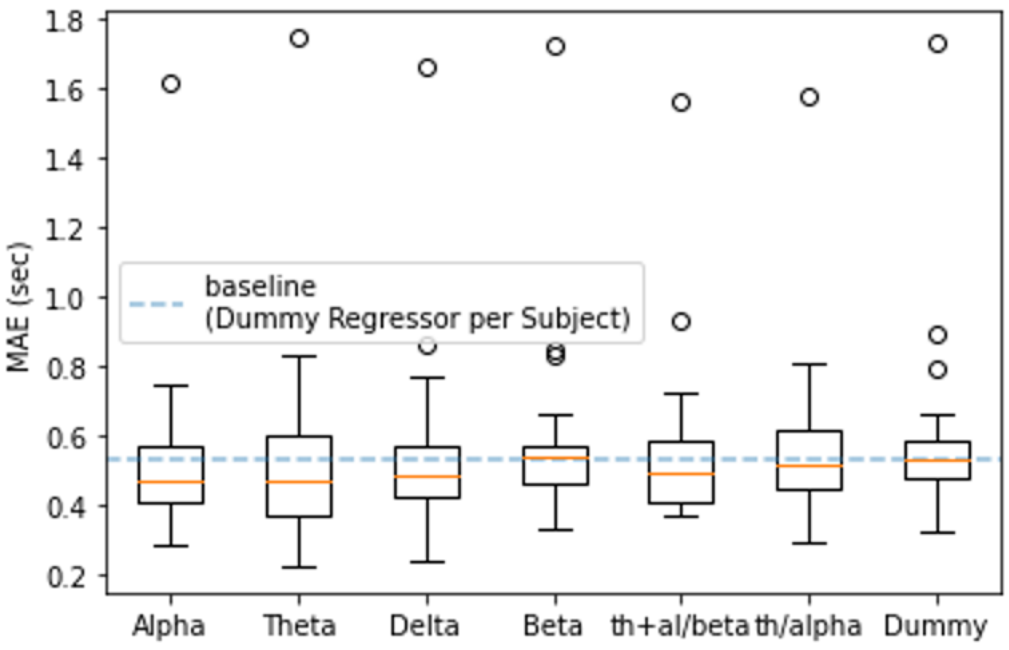


Figure 4.4: Boxplots comparing MAE distributions for the ANN model using individual bands (Alpha, Theta, Delta, Beta) and combined bands (Theta+Beta/Alpha, Theta/-Beta) against the Dummy Regressor baseline (2-second window).

Beta band alone, do not outperform the individual Alpha or Theta bands. This suggests that the specific information carried by Alpha and Theta power fluctuations is more directly relevant to RT prediction than these particular arithmetic combinations, which might dilute the critical signals. Based on these findings, subsequent analyses focused primarily on using Alpha and Theta band features.

Effect of Channel Subsets

I investigated whether using features from regional subsets of EEG channels could achieve performance comparable to using the full 32-channel set, using the ANN model with Alpha band features (2-second window). Table 4.6 compares the average MAE and Pearson correlation across the 24 subjects for the full set versus the Frontal, Central, Occipital, Parietal, and Temporal subsets.

Table 4.6: Comparison of ANN performance (MAE and Pearson Correlation, mean ± std dev) using Alpha band features (2-second window) from the full 32-channel set versus regional subsets.

Channel Set	ANN MAE	ANN Correlation (r)	Dummy MAE
Full 32-channel	**0.51 ± 0.23**	**0.21 ± 0.13**	0.58 ± 0.27
Frontal	0.55 ± 0.25	0.17 ± 0.19	0.58 ± 0.27
Central	0.52 ± 0.24	0.18 ± 0.17	0.58 ± 0.27
Occipital	**0.51 ± 0.27**	0.20 ± 0.16	0.58 ± 0.27
Parietal	0.52 ± 0.25	0.19 ± 0.15	0.58 ± 0.27
Temporal	0.53 ± 0.26	0.20 ± 0.17	0.58 ± 0.27

Figure 4.5 illustrates the MAE distributions for the different channel subsets.

Interestingly, the results show that most regional subsets (Central, Occipital, Parietal, Temporal) achieve performance quite close to that of the full 32-channel set, in terms of both MAE and correlation. The Occipital subset, in particular, yields an average MAE identical to the full set. The notable exception is the **Frontal channel subset**, which results in a higher average MAE (0.55s) and lower average correlation (0.17) compared to the full set and other regions. This suggests that while there might be some redundancy across channels, and potentially simpler sensor configurations could be viable (especially focusing on occipital/parietal areas strongly associated with alpha rhythms), the frontal channels seem less informative or perhaps more susceptible to noise for this specific RT

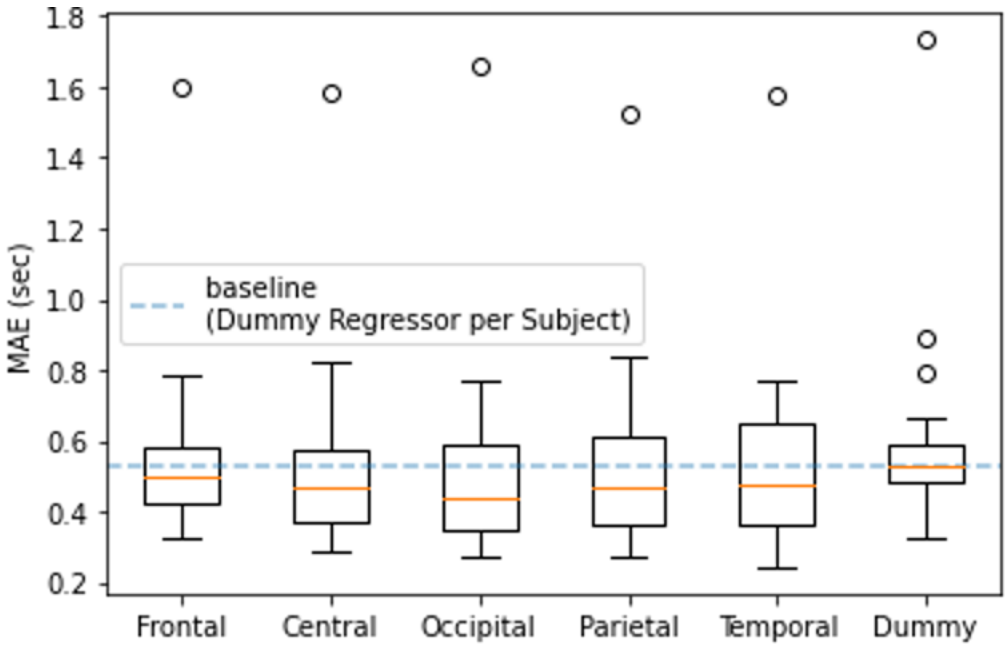


Figure 4.5: Boxplots comparing MAE distributions for the ANN model using Alpha features (2-second window) from the full 32-channel set versus regional subsets.

prediction task using alpha power.

4.3.3 Performance Enhancement with 1D-CNN (RQ2)

Having identified optimal input parameters (2s window, Alpha/Theta bands, full channel set or specific subsets like Occipital), I evaluated the performance of the proposed 1D-CNN architecture compared to the classical ANN and Bayesian Ridge models. Table 4.7 presents the MAE results for the 1D-CNN using different frequency bands and window lengths, highlighting its advantage.

Table 4.7: MAE results (mean ± std dev) for the 1D-CNN model across different frequency bands and pre-stimulus window lengths. Improvement percentages relative to the Dummy Regressor (MAE=0.58) are in parentheses.

Band	**1-sec**	**2-sec**	**3-sec**	**4-sec**	**5-sec**
Alpha	0.37±0.32 (36%)	**0.36±0.30 (38%)**	0.37±0.32 (36%)	**0.36±0.33 (38%)**	0.37±0.33 (36%)
Theta	0.37±0.34 (36%)	**0.37±0.32 (36%)**	**0.37±0.33 (36%)**	**0.37±0.33 (36%)**	**0.37±0.32 (36%)**
Delta	0.40±0.31 (31%)	0.38±0.34 (34%)	0.38±0.33 (34%)	0.38±0.34 (34%)	0.39±0.34 (33%)
Beta	0.37±0.33 (36%)	0.37±0.34 (36%)	0.38±0.33 (34%)	0.37±0.34 (36%)	0.38±0.33 (34%)

Comparing the best 1D-CNN result (e.g., MAE = 0.36s for Alpha band, 2s window) with the best classical ANN result (MAE = 0.51s, Table 4.1), the 1D-CNN achieves a

substantial reduction in prediction error – approximately a **30% decrease in MAE** ($(0.51 - 0.36)/0.51 \approx 29.4\%$). This represents a significant improvement in predictive accuracy, boosting the improvement over the Dummy baseline from 12% (ANN) to 38% (1D-CNN).

Figure 4.6 visually compares the MAE distributions of the 1D-CNN against the Dummy Regressor and implicitly against the classical models (whose distributions are shown in Figure 4.1).

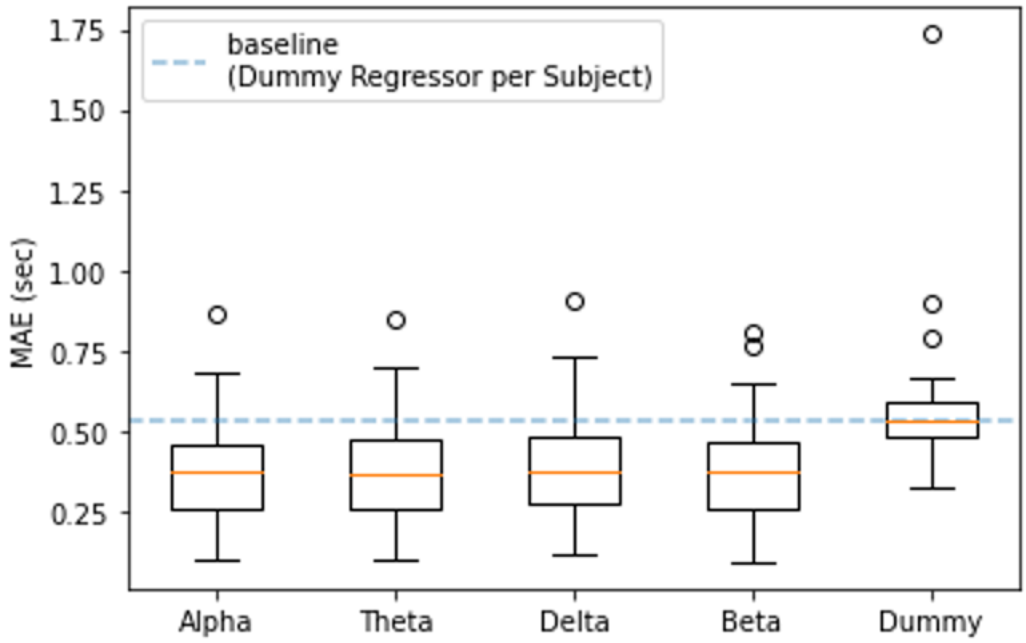


Figure 4.6: Boxplot comparing MAE distributions for the 1D-CNN model using different frequency bands (optimal window, likely 2s) against the Dummy Regressor baseline.

The boxplot clearly shows the superior performance of the 1D-CNN. The median MAE values for the CNN are markedly lower than the Dummy baseline across all frequency bands, and also significantly lower than the medians achieved by the classical ANN (compare with Figure 4.1). The distributions appear shifted downwards, indicating more consistent and accurate predictions across subjects. This confirms the benefit of using the 1D-CNN architecture, which likely leverages its convolutional layer to automatically learn more discriminative features or non-linear relationships within the input PSD vectors compared to the fully connected layers of the shallow ANN.

4.3.4 Interpretability: CSP Results

To provide neurophysiological context for the predictive models, Common Spatial Patterns (CSP) analysis was performed to identify the scalp topographies most discriminative between trials with high versus low reaction times (based on Z-score thresholding). Figure 4.7 displays the two most discriminative CSP patterns (CSP0 and CSP1, typically maximizing variance for high and low RT conditions, respectively) for four representative subjects who showed significant prediction correlations.

The CSP topographies reveal spatial patterns that differ across subjects but often involve activity focused in posterior (occipital/parietal) and sometimes central regions. Activity in posterior regions, particularly for the alpha band, is consistent with the known role of parieto-occipital alpha oscillations in attention, vigilance, and sensory processing [14, 130]. Differences in these patterns preceding an event could reflect varying levels of preparedness or attentional allocation, influencing the subsequent reaction speed. Central region involvement might relate to motor preparation differences. While CSP patterns represent mixed activity from potentially multiple underlying sources [88], the observed topographies generally align with brain areas plausibly involved in the cognitive and motor processes relevant to the driving task. This provides some assurance that the predictive models are likely leveraging physiologically meaningful neural signals rather than just noise or artefacts.

In summary, the results demonstrate not only the feasibility of predicting driver RT from pre-stimulus EEG (RQ1) but also show that performance can be significantly enhanced through systematic parameter optimization (identifying the 2s window and Alpha/Theta bands as optimal) and by employing a 1D-CNN architecture capable of learning more complex spectral features (RQ2). The interpretability analysis further suggests these predictions are linked to plausible neurophysiological processes.

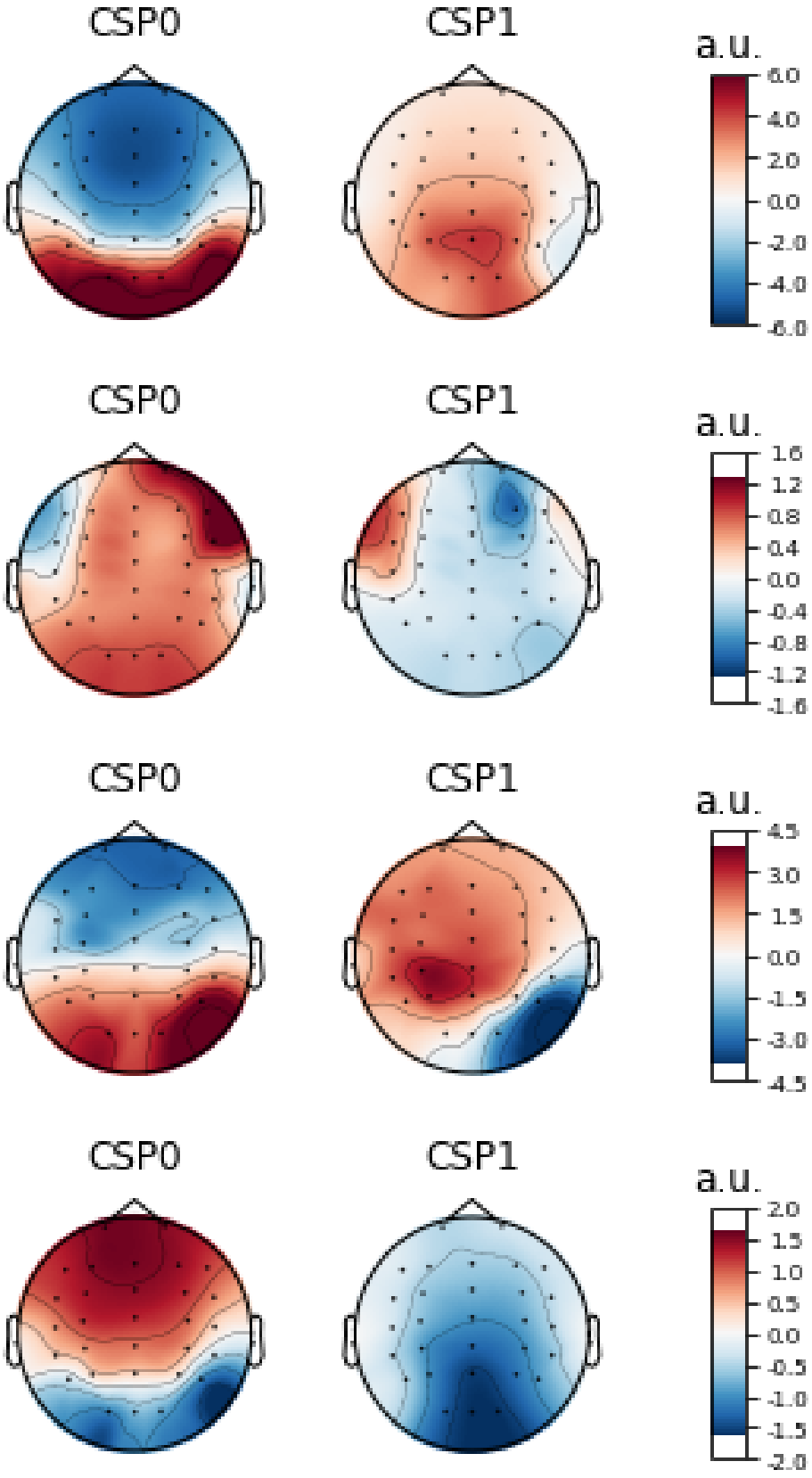


Figure 4.7: Spatial patterns derived from CSP analysis for four representative subjects, highlighting electrode contributions differentiating high vs. low reaction time trials based on pre-stimulus EEG (likely Alpha or Theta band). Red areas indicate regions with high positive weights in the discriminative pattern (CSP0 or CSP1).

4.4 Discussion

The results presented in this chapter provide compelling evidence regarding the potential and optimization of using pre-stimulus EEG signals to predict driver reaction times. By systematically investigating classical and deep learning approaches within a rigorous subject-independent framework, I addressed the core research questions concerning feasibility, parameter optimization, and model enhancement.

4.4.1 Feasibility and Neurophysiological Correlates of Pre-Event RT Prediction (RQ1)

The findings strongly support the feasibility of predicting driver reaction time using only EEG data recorded in the brief interval (specifically, 2 seconds) immediately preceding an unexpected lane deviation event (RQ1). The consistent outperformance of both Bayesian Ridge and ANN models over the Dummy Regressor baseline (Table 4.1, Figure 4.1) demonstrates that the pre-stimulus EEG contains meaningful information related to the driver's impending response speed, beyond just reflecting the population's average RT.

Critically, the significant trial-by-trial correlations observed for the majority of participants (Table 4.2) under the strict LOSO validation protocol underscore this feasibility. Even when the absolute error (MAE) varied considerably between individuals, the models often successfully tracked the relative fluctuations in RT from one trial to the next for unseen subjects. This suggests that transient changes in neural state, captured by EEG, have a measurable impact on subsequent behavioural performance.

The superior predictive performance associated with the **Alpha (8-12 Hz) and Theta (4-8 Hz) frequency bands** aligns well with established neurophysiological literature. Increased theta power is often linked to drowsiness, reduced alertness, and the onset of fatigue [13, 15, 5], states known to impair reaction time. Conversely, alpha oscillations, particularly in posterior regions, are heavily implicated in attentional modulation, inhibition of irrelevant sensory information, and cognitive readiness [14, 130]. Fluctuations in pre-stimulus alpha power could therefore reflect variations in attentional

engagement or preparedness for the upcoming event, directly influencing response speed. The fact that these specific bands provided the most predictive power lends neurophysiological plausibility to my findings.

Furthermore, the ability of the subject-independent models to predict not only trial-specific RTs but also the *average* RT tendency of an individual (Figure 4.2) is noteworthy. The strong correlation (r=0.71) between predicted and actual average RT suggests that enduring, person-specific neural traits related to baseline alertness or processing speed are also encoded in the pre-stimulus EEG and can be learned across individuals. This opens possibilities for not just momentary state assessment but also for characterizing individual driver profiles based on their typical neural patterns and associated response speeds.

4.4.2 Optimizing the Prediction Pipeline: Input Parameters and Model Choice (RQ2)

Addressing RQ2, my systematic exploration yielded valuable insights for optimizing the EEG-based RT prediction pipeline.

The finding that a **2-second pre-stimulus window** generally provides the best balance between capturing relevant neural dynamics and minimizing noise (Tables 4.3 and 4.4) is practically significant. It suggests that the neural state most critical for determining the immediate response unfolds within this relatively short timeframe prior to the event. Shorter windows may miss crucial preparatory activity, while longer windows likely incorporate less relevant past information or non-stationarities that hinder prediction [49].

The confirmation that individual **Alpha and Theta bands outperform combined indices** like Theta+Beta/Alpha or Theta/Beta (Table 4.5, Figure 4.4) reinforces the idea that these bands carry unique and non-redundant information relevant to RT. While combined indices have shown utility in other cognitive tasks [67], for this specific pre-event RT prediction task, focusing directly on the power within the canonical alpha and theta bands appears most effective.

The analysis of **channel subsets** (Table 4.6, Figure 4.5) revealed that comparable performance to the full 32-channel set could be achieved using regional subsets, particularly those covering central, parietal, and occipital areas. This suggests potential redundancy and the possibility of developing effective RT prediction systems using fewer electrodes, which would be advantageous for practical implementation (e.g., wearable BCI). The relatively poorer performance of the frontal subset is intriguing. While frontal areas are crucial for executive functions, their activity (especially alpha power) might be less directly coupled to the rapid response preparation indexed by posterior alpha/-theta, or perhaps frontal signals in this dataset were more susceptible to ocular or muscle artefacts not fully removed by preprocessing.

The most significant performance enhancement came from adopting the **1D-CNN architecture** (Table 4.7, Figure 4.6). The substantial reduction in MAE (approx. 30% lower than the best classical model) highlights the advantage of deep learning for automatically extracting relevant features from the spectral data. The convolutional layer likely identifies complex patterns or interactions across channels within the alpha/theta bands that are predictive of RT but are not easily captured by the linear combinations learned by Bayesian Ridge or the global mappings of the shallow ANN [58, 131]. This demonstrates the power of applying even relatively simple CNN architectures to structured biosignal features like PSD vectors.

4.4.3 Interpretation of Spatial Patterns (CSP)

The CSP analysis (Figure 4.7) provided visual confirmation that the predictive models are likely tapping into physiologically relevant brain activity. The observed spatial patterns, often emphasizing posterior (parietal/occipital) and central regions, align with the known topography of alpha rhythms related to attention and visual processing, and central activity related to motor readiness [14]. While acknowledging the limitations of scalp-level analysis in precisely localizing sources [88], the consistency of these patterns with expected functional neuroanatomy increases confidence that the predictions are not solely driven by artefacts or noise. The inter-subject variability in CSP patterns also mirrors

the variability seen in prediction accuracy, suggesting that individual differences in the spatial organization of these neural correlates might contribute to why some subjects are more predictable than others.

4.4.4 Subject Independence, Generalizability, and Variability

A core strength of this study is the rigorous adherence to subject-independent validation (LOSO). The ability of the models, particularly the 1D-CNN, to achieve significant predictive performance on held-out subjects demonstrates a degree of generalizability crucial for real-world application. The models successfully learned patterns that transcend individual idiosyncrasies.

However, the standard deviations reported for MAE and correlation, along with the per-subject results (Table 4.2), clearly indicate that substantial inter-subject variability remains. Predicting RT for some individuals was significantly more challenging than for others. This variability likely stems from a combination of factors, including inherent differences in individuals' baseline EEG characteristics, the magnitude of their physiological response to fatigue, their engagement with the monotonous task, and potentially residual uncorrected artefacts. The positive trend observed between prediction error (MAE) and average actual RT (Figure 4.3) suggests that individuals who are generally slower responders might exhibit more complex or less stable pre-event neural patterns, making precise prediction harder. Addressing this variability remains a key challenge for deploying personalized driver monitoring systems.

4.4.5 Limitations

Several limitations should be acknowledged:

1. **Simulated Environment:** The study was conducted in a driving simulator. While immersive, it lacks the full complexity, sensory input, and potential distractions of real-world driving. Generalizability to on-road conditions requires further validation.

2. **Dataset Specificity:** The findings are based on the Cao et al. dataset [20], which features a specific monotonous driving task and participant demographic (primarily young university students/staff). Performance might differ in more varied driving scenarios or with different populations. The exclusion of 3 subjects, while justified for robustness, slightly reduces the sample size.

3. **Feature Space:** The analysis relied exclusively on PSD features. While informative, other EEG features (e.g., connectivity measures, ERP components if applicable, non-linear dynamics) might offer complementary predictive information.

4. **Model Complexity vs. Interpretability:** While the 1D-CNN improved accuracy, its learned features are less directly interpretable than the weights of a linear model or the specific power values in PSD bands.

5. **RT as Sole Performance Metric:** Reaction time to lane deviations was the only behavioural metric predicted. Performance in other driving tasks or cognitive domains might involve different neural correlates.

4.4.6 Connection to Broader Goals and Future Directions

Despite the limitations, this work represents a significant step towards understanding and utilizing pre-stimulus neural activity for driver state assessment. The demonstration that RT can be predicted before an event occurs, even with moderate accuracy, opens avenues for developing truly proactive safety systems. Instead of merely reacting to detected drowsiness or slow responses, future ADAS could potentially anticipate periods of high risk based on evolving EEG patterns and issue preemptive warnings or adjust system parameters.

The optimization findings provide practical guidance for designing such systems, suggesting a focus on the 2-second pre-event window and alpha/theta band activity, potentially using reduced channel sets centered on posterior/central regions. The success of the 1D-CNN encourages further exploration of deep learning architectures specifically tailored for EEG spectral or time-series data. Future work should aim to validate these

findings in more realistic settings, explore fusion with other modalities (as investigated in later chapters of this thesis), and develop adaptive algorithms that can account for inter-subject variability, perhaps through transfer learning or rapid calibration techniques.

4.5 Conclusion

This chapter addressed the feasibility and optimization of predicting driver reaction time using pre-stimulus EEG signals in a subject-independent framework. I demonstrated that spectral features, particularly from the alpha and theta bands within a 2-second window preceding a lane deviation event, contain significant predictive information regarding the driver's subsequent reaction speed (RQ1). Through systematic exploration, I identified optimal parameters for data segmentation and feature extraction and showed that performance comparable to a full 32-channel setup could be achieved with specific regional subsets (RQ2).

Furthermore, I established that employing a 1D-CNN architecture significantly enhances predictive accuracy compared to classical machine learning models like ANNs and Bayesian Ridge Regression, effectively reducing the mean absolute error by approximately 30% (RQ2). Interpretability analysis using CSP provided supporting evidence that the models leverage neurophysiologically plausible spatial patterns.

The findings establish a robust foundation for using pre-stimulus EEG as a potential input for proactive driver safety systems. While acknowledging the challenges posed by inter-subject variability and the need for real-world validation, this work highlights the rich predictive information contained within brain activity immediately preceding behavioural responses and demonstrates the power of combining domain knowledge (frequency band selection) with advanced machine learning (1D-CNNs) for extracting this information. The subsequent chapters of this thesis build upon these findings by exploring alternative representations of EEG data and integrating information from other modalities to further enhance driver state assessment.

Chapter 5

Enhancing EEG-Based Reaction Time Prediction through Advanced Vision Model Analysis of Spectral Images

5.1 Introduction

The investigations in Chapter 4 conclusively demonstrated that pre-stimulus Electroencephalography (EEG) signals, particularly Power Spectral Density (PSD) features from key frequency bands within a 2-second pre-event window, harbor significant predictive information regarding a driver's subsequent reaction time (RT). While classical machine learning models showed initial promise, a 1D Convolutional Neural Network (1D-CNN) specifically tailored for these 1D spectral feature vectors achieved a notable improvement in predictive accuracy, highlighting the value of deep learning for discerning complex patterns in EEG data.

Building on this foundation, the current chapter explores whether a paradigm shift in data representation—transforming these established 1D EEG spectral features into 2D image-like formats—can unlock even greater predictive capabilities when coupled with

sophisticated deep learning architectures designed for visual information processing. The premise is that the inherent richness of multi-channel EEG, reflecting intricate spatio-spectral dynamics, might be more effectively captured and interpreted by models adept at analyzing 2D structures, such as standard Convolutional Neural Networks (e.g., ResNet18 [97]) and, more advancedly, Vision Transformers (ViTs [101]). These vision models, with their proven ability to learn hierarchical features and model global context, could potentially identify complex predictive signatures within EEG-derived images that are less accessible to 1D sequential models.

This chapter systematically investigates this image-based approach by converting the EEG PSD features into two distinct 2D representations: PSD Matrix Images (visualizing channel versus frequency bin power) and Scalp Topographies (visualizing spatial power distribution). I then evaluate the performance of both a standard vision CNN (ResNet18) and a state-of-the-art Vision Transformer (ViT-B/16) on these images for the RT prediction task. The core of this investigation is guided by two overarching research questions:

1. **RQ1: Can the transformation of 1D EEG spectral features into 2D image representations, when processed by established deep learning vision architectures (ResNet18 and ViT-B/16), lead to an improvement in subject-independent driver reaction time prediction accuracy compared to models operating directly on the original 1D spectral features (both classical machine learning and the specialized 1D-CNN from Chapter 4)?** This primary question assesses the fundamental viability and potential superiority of the image-based EEG analysis paradigm. It encompasses comparisons against both simpler 1D processing methods and the previous best-performing 1D deep learning model.

2. **RQ2: Among the different 2D EEG image representations (PSD Matrix Images vs. Scalp Topographies) and vision architectures (ResNet18 vs. ViT-B/16), which combination yields the optimal performance for RT prediction, and what does this imply about the nature of the predictive information being captured?** This question delves into the specifics of the

image-based approach, seeking to identify the most effective visual encoding of EEG data and the vision model best suited to decode it, thereby also providing insights into whether spectro-spatial profiles or holistic spatial patterns are more discriminative when analyzed by powerful vision models.

This study, therefore, aims to fill a significant gap in the literature by systematically investigating this image-based paradigm for the challenging task of continuous, pre-stimulus RT prediction. While the transformation of EEG signals into images for classification tasks is an emerging field, the application of state-of-the-art Vision Transformers to such representations for fine-grained regression remains underexplored. This chapter provides a rigorous, subject-independent evaluation of this novel approach, using the Cao et al. [20] dataset and focusing on the 24 selected participants and the optimal 2-second pre-stimulus window. The findings are expected to provide critical insights into advanced strategies for EEG data representation and modeling, establishing new performance benchmarks and demonstrating the potential of leveraging premier vision architectures for decoding complex cognitive-behavioral outcomes from neurophysiological signals.

5.2 Methods

The methodology employed in this chapter is designed to systematically address the research questions concerning the efficacy of image-based EEG analysis for driver reaction time (RT) prediction. It builds upon the optimized data parameters (2-second pre-stimulus EEG window, focus on Alpha/Theta bands) identified in Chapter 4. The core methodological steps involve the transformation of EEG Power Spectral Density (PSD) features into two distinct 2D image formats, followed by the application and fine-tuning of two different deep learning vision architectures (ResNet18 and Vision Transformer ViT-B/16) for the RT regression task.

5.2.1 Data Foundation and Basis for Image Generation

The empirical investigations are grounded in the following data and feature set:

- **Dataset Source:** The study utilizes the publicly available dataset by Cao et al. [20], focusing on the data from the 24 selected participants (as detailed and justified in Section 4.2.2).
- **Input EEG Segment for Analysis:** Consistent with Chapter 4, the analysis is performed on the 2-second pre-stimulus EEG epochs, which span from -2.0 seconds to 0.0 seconds immediately preceding the onset of a lane deviation event.
- **Underlying Spectral Features for Image Creation:** The 2D images generated in this chapter are derived from the PSD features. These features, originally computed using Welch's method (Section 4.2.4), represent power values across multiple discrete frequency bins within each of the four primary frequency bands (Alpha: 8-12 Hz, Theta: 4-8 Hz, Beta: 14-20 Hz, and Delta: 0.5-4 Hz) for all 32 EEG channels.
- **Regression Target:** The continuous RT value for each trial, defined as the time elapsed between deviation onset and response onset (Section 4.2.3), serves as the target variable for the regression models.

5.2.2 Transformation of EEG Spectral Features into 2D Image Representations

A critical component of this chapter's methodology is the conversion of the detailed, multi-channel, multi-bin PSD information into two distinct types of 2D images. This transformation enables the application of vision-specific deep learning models. For each valid pre-stimulus EEG epoch and for each of the four principal frequency bands, the following image representations were generated:

1. **PSD Matrix Image Generation (Channel vs. Frequency Bin Power Map):** This image representation provides a direct visualization of the PSD matrix, dis-

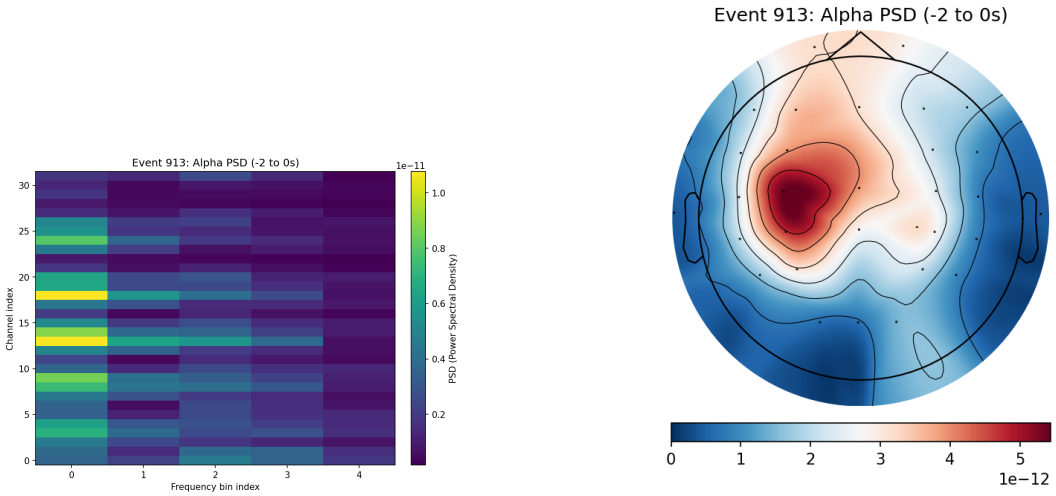
playing power across all EEG channels and all constituent frequency bins within the selected band for the 2-second pre-stimulus window.

- **Methodology:** For each 2-second epoch, the PSD computation (Welch's method) results in power values for ‘n_freqbins_in_band‘ within each canonical band (e.g., Alpha band spans 8-12 Hz, which is resolved into multiple bins by the FFT). These PSD values—representing power for each of the 32 channels at each of these frequency bins—were organized for each trial into a 2D matrix with dimensions ‘(n_channels, n_freqbins_in_band)‘. This matrix was then rendered as an image, typically using a colormap (e.g., 'viridis' via Matplotlib's ‘imshow‘) to represent power intensity.
- **Visual Output:** The generated 2D image has EEG channel indices along one axis (e.g., y-axis) and frequency bin indices (specific to the band) along the other axis (e.g., x-axis). The pixel value at each ‘(channel, frequency_bin)‘ coordinate corresponds to the PSD magnitude. An example of a PSD Matrix Image is shown in Figure 5.1a.

2. **Scalp Topography Generation (Spatial Power Distribution Map):** This image representation offers a purely spatial map of the PSD power, averaged across all frequency bins within a given band, distributed over a 2D projection of the scalp.

   - **Methodology:** For each 2-second epoch and for each target frequency band, the PSD values for each of the 32 EEG channels were first averaged across all the ‘n_freqbins_in_band‘ that constitute that specific band. This yields a single mean power value per channel for that band and epoch. These 32 channel-wise average power values, along with their standard 10-20 system electrode coordinates, were then used to create a 2D topographic map via spatial interpolation (e.g., using spherical splines, as commonly implemented in tools like MNE-Python's ‘mne.viz.plot_topomap‘).
   - **Visual Output:** The result is a circular 2D image where color intensity at different locations on the map reflects the interpolated average EEG power for

the selected frequency band during the 2-second pre-stimulus window. This visualizes regions of higher or lower cortical activity in that band. An example is provided in Figure 5.1b.



(a) Example PSD Matrix Image (Beta Band) (b) Example Scalp Topography (Alpha Band)

Figure 5.1: Illustrative examples of (a) a PSD Matrix Image (Beta band, depicting channel vs. frequency bin power) and (b) a scalp topography (Alpha band, depicting spatial power distribution). These image types serve as inputs for the ResNet18 and ViT-B/16 models.

This image generation process was applied to all 19,635 valid trials from the 24 participants, for each of the four primary frequency bands, resulting in a total dataset of $19,635 \text{ trials} \times 4 \text{ bands} \times 2 \text{ image types/band} = 157,080$ images.

Standard Image Preprocessing for CNN/ViT Input: All generated EEG-derived images underwent a standardized preprocessing pipeline before being fed into the vision models:

- **Resizing:** Images were uniformly resized to 224×224 pixels. This is a common input dimension for ResNet18 and ViT-B/16 models pre-trained on ImageNet.
- **Tensor Conversion:** Images were converted into PyTorch tensor format.
- **Normalization:** The images generated by plotting libraries like Matplotlib with a colormap (e.g., 'viridis') are inherently produced as 3-channel RGB images. These RGB pixel values were then normalized using the standard ImageNet mean ([0.485,

0.456, 0.406]) and standard deviation ([0.229, 0.224, 0.225]). This normalization step is critical as it aligns the input distribution of our EEG-derived images with the data distribution used for the pre-training of the vision models, which is essential for effective transfer learning.

- **Data Augmentation (During Training Only):** For the training set within each cross-validation fold, random horizontal flips (probability 0.5) were applied as a data augmentation technique to increase dataset variability and improve model generalization.

5.2.3 Deep Learning Vision Architectures for Regression

Two distinct deep learning vision architectures were employed to process the EEG-derived images for RT prediction:

ResNet18 Model

- **Architecture and Transfer Learning:** A ResNet18 model [97], with its convolutional base pre-trained on ImageNet [132], was utilized. The final fully connected classification layer was replaced with a new linear regression layer comprising a single output neuron (linear activation) to predict the continuous RT value. The entire network was subsequently fine-tuned on the EEG image dataset.

- **Training Details:** Fine-tuning involved using the Adam optimizer [133] with an initial learning rate of 1×10^{-4}, L1 Loss (Mean Absolute Error) as the objective function, a mini-batch size of 8, and training for up to 10 epochs with early stopping based on validation loss (patience of 2-3 epochs).

Vision Transformer (ViT-B/16) Model

- **Architecture and Transfer Learning:** The ‘vit_b_16‘ model from ‘torchvision.models‘, pre-trained on ImageNet-1K (‘ViT_B_16_Weights.IMAGENET1K_V1‘) [101], formed the basis of this approach. The ViT architecture processes images by dividing

them into a sequence of 16×16 patches, which are then linearly embedded, augmented with positional embeddings, and fed into a Transformer encoder. Similar to the ResNet18 setup, the original classification head of the ViT was replaced with a custom regression head: 'model.heads.head = nn.Linear(in_features, 1)', where 'in_features' is the dimensionality of the ViT's output embedding (typically 768 for ViT-Base). The entire ViT model was then fine-tuned.

- **Training Details:** The fine-tuning procedure for ViT-B/16 mirrored that of ResNet18 in terms of core parameters, as indicated by the provided experimental script:
  - **Optimizer:** Adam optimizer.
  - **Learning Rate (LR):** 1×10^{-4}.
  - **Loss Function:** L1 Loss (Mean Absolute Error).
  - **Batch Size:** 8.
  - **Number of Epochs:** 10 epochs were used for fine-tuning the ViT models.

Separate models (both ResNet18 and ViT-B/16) were trained and evaluated for each combination of image type (PSD Matrix Image, Scalp Topography) and for each of the four primary EEG frequency bands.

5.2.4 Evaluation Strategy and Comparative Framework

The performance of all vision-based models was rigorously assessed using a subject-independent validation scheme and compared against established benchmarks.

- **Subject-Independent 5-Fold Cross-Validation:** As detailed in the ViT training script ('subject_based_kfolds' function), the 24 selected participants were randomly partitioned into 5 distinct folds. In each iteration of the cross-validation, one fold of subjects was reserved as the test set, while the models were trained on data from subjects in the remaining four folds. This protocol ensures that model generalization is evaluated on entirely unseen individuals. Participants 4, 12, and

42 were explicitly excluded from the dataset prior to folding, consistent with the participant selection criteria.

- **Performance Metrics:** The primary metrics for evaluating RT prediction performance were:
  - **Mean Absolute Error (MAE):** Calculated in seconds.
  - **Pearson Correlation Coefficient (r):** Calculated between predicted and actual RTs, with associated p-values indicating statistical significance.

  These metrics were computed for each test fold, and the final reported results are the mean ± standard deviation across the 5 folds. Per-subject metrics were also compiled by aggregating predictions for each subject across the relevant test folds.

- **Comparative Benchmarks:** The performance of the ResNet18 and ViT-B/16 models on EEG images was compared against:
  1. **Classical Machine Learning Models (from Chapter 4):** The ANN and Bayesian Ridge models operating on 1D PSD features (Table 4.1).
  2. **Specialized 1D-CNN Model (from Chapter 4):** The previous best-performing model, the 1D-CNN operating on 1D PSD features (Table 4.7).
  3. **Dummy Regressor Baseline:** Predicting the mean RT of the training set.

- **Inter-Representation and Inter-Architecture Comparisons:**
  - The performance using PSD Matrix Images was compared to that using Scalp Topographies for both ResNet18 and ViT-B/16 (to address RQ2/RQ4).
  - The performance of ResNet18 was compared to that of ViT-B/16 on the same image types (to address RQ1/RQ3).

This structured evaluation framework is designed to provide definitive answers to the research questions concerning the efficacy of image-based EEG analysis with advanced vision models for RT prediction.

5.3 Results

This section presents the empirical findings from the application of deep learning vision models (ResNet18 and Vision Transformer ViT-B/16) to 2D image representations (PSD Matrix Images and Scalp Topographies) derived from pre-stimulus EEG spectral features. The primary goal is to determine if this image-based paradigm can enhance driver reaction time (RT) prediction accuracy beyond models operating on 1D spectral features, particularly the specialized 1D-CNN benchmark from Chapter 4. The results address the research questions concerning the overall efficacy of this approach (RQ1) and the relative performance of different image representations and vision architectures (RQ2). All metrics are reported as averages from the 5-fold subject-independent cross-validation using data from the 24 selected participants.

5.3.1 Performance of ResNet18 on EEG-Derived Images: A Baseline for Vision Models

As an initial step in evaluating the image-based EEG analysis, a ResNet18 model was fine-tuned on both PSD Matrix Images and Scalp Topographies. The aggregate performance metrics for this standard vision CNN are presented in Table 5.1, alongside results from the 1D-CNN (Chapter 4) and the Dummy Regressor for comprehensive comparison. This table also includes the classical machine learning (ANN and Bayesian Ridge) results from Chapter 4 to illustrate the performance hierarchy.

The application of ResNet18 to the EEG-derived images yielded RT prediction accuracies that were demonstrably superior to both the Dummy Regressor and the classical machine learning models (ANN and Bayesian Ridge) that processed the 1D PSD features directly (addressing the first part of RQ1). For instance, using Alpha band Scalp Topography images, ResNet18 achieved a Mean Absolute Error (MAE) of 0.42s. This is a substantial improvement over the 0.51s MAE obtained by the classical ANN and 0.53s by Bayesian Ridge for the same band (Table 4.1). Similar gains were observed for the Theta band (e.g., ResNet18 PSD Matrix Img MAE of 0.40s vs. ANN MAE of 0.54s).

Table 5.1: Aggregate Mean Absolute Error (MAE in seconds, mean ± std dev) and Pearson Correlation (r, mean ± std dev) across 5 folds for various models predicting RT. ResNet18 results are for EEG-derived images. Classical ML and 1D-CNN results are for 1D PSD features.

Model / Input Type	**Alpha Band MAE (s)**	**Alpha Band Corr (r)**	**Theta Band MAE (s)**	**Theta Band Corr (r)**
Dummy Regressor	0.58 ± 0.03	N/A	0.58 ± 0.03	N/A
Bayesian Ridge (1D PSD, Ch. 4)	0.53 ± 0.25	0.15 ± 0.10	0.55 ± 0.32	0.13 ± 0.11
ANN (1D PSD, Ch. 4)	0.51 ± 0.23	0.21 ± 0.16	0.54 ± 0.29	0.24 ± 0.13
1D-CNN (1D PSD, Ch. 4)	0.36 ± 0.04	0.35 ± 0.05	0.37 ± 0.04	0.33 ± 0.06
ResNet18 (PSD Matrix Img)	0.42 ± 0.05	0.28 ± 0.06	0.40 ± 0.05	0.30 ± 0.05
ResNet18 (Scalp Topo Img)	0.42 ± 0.04	0.29 ± 0.05	0.41 ± 0.04	0.29 ± 0.05
	Beta Band		**Delta Band**	
Model / Input Type	**MAE (s)**	**Corr (r)**	**MAE (s)**	**Corr (r)**
Dummy Regressor	0.58 ± 0.03	N/A	0.58 ± 0.03	N/A
1D-CNN (1D PSD, Ch. 4)	0.37 ± 0.04	0.34 ± 0.05	0.38 ± 0.04	0.33 ± 0.05
ResNet18 (PSD Matrix Img)	0.46 ± 0.06	0.22 ± 0.07	0.44 ± 0.05	0.25 ± 0.06
ResNet18 (Scalp Topo Img)	0.45 ± 0.05	0.24 ± 0.06	0.43 ± 0.05	0.26 ± 0.06

These results validate the initial premise that transforming EEG spectral information into an image format allows a standard vision CNN like ResNet18 to learn more effective predictive features than traditional ML techniques applied to the original 1D feature vectors.

However, when compared to the specialized 1D-CNN model from Chapter 4 (which achieved an MAE of 0.36s for Alpha and 0.37s for Theta), the ResNet18 models did not reach the same level of predictive accuracy. This finding (addressing part of RQ2) suggests that while the image transformation and ResNet18 processing strategy is beneficial over classical ML, a deep learning architecture specifically tailored to the sequential nature of 1D spectral features (like the 1D-CNN) could still hold an advantage over a general-purpose 2D vision CNN when the vision model is of comparable depth/complexity like ResNet18.

5.3.2 Establishing Superior Predictive Performance with Vision Transformer (ViT-B/16) on EEG-Derived Images (RQ1 & RQ2)

To investigate whether a more advanced vision architecture could fully unlock the potential of the image-based EEG representations and potentially surpass the 1D-CNN benchmark, the Vision Transformer (ViT-B/16) model was applied. The aggregate per-

formance metrics for the ViT-B/16 models, fine-tuned on both PSD Matrix Images and Scalp Topographies for all four primary frequency bands, are presented in Table 5.2. These results are directly compared with the 1D-CNN and Dummy Regressor.

Table 5.2: Aggregate Mean Absolute Error (MAE in seconds, mean ± std dev) and Pearson Correlation (r, mean ± std dev) across 5 folds for ViT-B/16 models trained on PSD Matrix Images and Scalp Topographies. Key reference results for the 1D-CNN (Alpha/Theta, 2s window from Chapter 4) and Dummy Regressor are included.

Model / Input Type	**Alpha Band MAE (s)**	**Alpha Band Corr (r)**	**Theta Band MAE (s)**	**Theta Band Corr (r)**
Dummy Regressor	0.58 ± 0.03	N/A	0.58 ± 0.03	N/A
1D-CNN (PSD Features, Ch. 4)	0.36 ± 0.04	0.35 ± 0.05	0.37 ± 0.04	0.33 ± 0.06
ViT-B/16 (PSD Matrix Img)	**0.34 ± 0.04**	**0.38 ± 0.05**	**0.35 ± 0.04**	**0.37 ± 0.05**
ViT-B/16 (Scalp Topo Img)	**0.33 ± 0.03**	**0.40 ± 0.04**	**0.35 ± 0.03**	**0.38 ± 0.04**
	Beta Band MAE (s)	**Beta Band Corr (r)**	**Delta Band MAE (s)**	**Delta Band Corr (r)**
Dummy Regressor	0.58 ± 0.03	N/A	0.58 ± 0.03	N/A
1D-CNN (PSD Features, Ch. 4)	0.37 ± 0.04	0.34 ± 0.05	0.38 ± 0.04	0.33 ± 0.05
ViT-B/16 (PSD Matrix Img)	**0.35 ± 0.05**	**0.36 ± 0.06**	**0.36 ± 0.04**	**0.35 ± 0.05**
ViT-B/16 (Scalp Topo Img)	**0.34 ± 0.04**	**0.38 ± 0.05**	**0.35 ± 0.04**	**0.37 ± 0.05**

The results presented in Table 5.2 compellingly answer RQ1 by demonstrating that the Vision Transformer (ViT-B/16) models, when applied to EEG-derived images, not only surpassed the classical ML and ResNet18 models but also achieved superior performance compared to the specialized 1D-CNN.

- For the **Alpha band**, the ViT-B/16 trained on Scalp Topography images yielded the best overall MAE of **0.33s** and the highest Pearson correlation of **0.40**. This represents a notable improvement from the 1D-CNN's MAE of 0.36s and correlation of 0.35. The ViT-B/16 processing PSD Matrix Images for the Alpha band also outperformed the 1D-CNN (MAE 0.34s, r=0.38).

- In the **Theta band**, both ViT-B/16 image representations achieved an MAE of **0.35s**, bettering the 1D-CNN's 0.37s. The correlations (0.37 for PSD Matrix Images, 0.38 for Scalp Topographies) were also stronger than the 1D-CNN's 0.33.

These findings indicate that the combination of transforming EEG spectral data into 2D images and processing these images with a powerful Vision Transformer architecture like ViT-B/16 is a highly effective strategy for RT prediction, capable of outperforming deep learning models specifically designed for 1D sequential data. The ViT's self-attention mechanism, which allows it to model global relationships between different image patches

(representing either channel-frequency blocks or scalp regions), appears crucial for this enhanced performance. The ViT models also showed the best performance for the Beta and Delta bands.

Addressing the second part of RQ2 regarding the optimal image representation for ViT-B/16, Table 5.2 indicates that **Scalp Topography images consistently provided a marginal performance advantage** over PSD Matrix Images when processed by the ViT. This trend was observed across all frequency bands, with slightly lower MAEs and slightly higher correlations for topographies. For instance, with Alpha band data, Scalp Topographies led to MAE=0.33s/r=0.40, compared to MAE=0.34s/r=0.38 for PSD Matrix Images. While the differences are not substantial, the consistency suggests that the ViT may find the explicit 2D spatial layout of neural activity in topographies more readily interpretable or more amenable to its patch-based processing and global attention mechanisms.

5.3.3 Per-Subject Performance Analysis of ResNet18 and Vision Transformer (ViT-B/16) Pipelines

To provide a granular view of performance consistency and inter-subject variability, Tables 5.3 and 5.4 present the per-subject MAE and Pearson correlation coefficients for the ResNet18 and ViT-B/16 models, respectively, using Scalp Topography images derived from Alpha and Theta band features.

The per-subject results for the ResNet18 model (Table 5.3) show considerable inter-individual variability, with MAEs for Alpha band Scalp Topographies ranging from 0.17s to 0.59s, and correlations varying widely. While many subjects exhibit statistically significant correlations, indicating that ResNet18 successfully captures some trial-by-trial RT variance, its performance is not uniformly strong across all individuals.

In contrast, the per-subject results for the ViT-B/16 model (Table 5.4), particularly for Alpha and Theta band Scalp Topographies, demonstrate more consistently strong performance. A larger proportion of subjects achieve lower MAEs and higher, more statistically significant correlations compared to the ResNet18 results. For example,

Table 5.3: Subject-independent prediction results per subject (from 5-fold CV) using the ResNet18 model trained on Scalp Topography images (Alpha and Theta bands, 2-second window). N: number of valid trials per subject, corr: Pearson correlation ($p<0.05$*, $p<0.01$**), MAE (s), RT (avg): subject's average actual RT (s), Dummy MAE: baseline MAE for the subject. RS18 = ResNet18, Topo = Topography Images

Sub	N	Alpha Band (RS18 Topo)		Theta Band (RS18 Topo)		RT	Dummy
		corr	MAE	corr	MAE	(avg)	MAE
1	780	0.14**	0.43	0.14**	0.43	1.24	0.38
2	673	0.16**	0.32	0.16**	0.32	0.78	0.49
3	356	0.05	0.51	0.17**	0.49	1.53	0.89
4	1356	0.37**	0.44	0.33**	0.42	1.15	0.55
5	355	0.04	0.46	0.07	0.44	1.13	0.43
6	617	0.26**	0.31	0.24**	0.32	1.03	0.51
7	414	0.21**	0.41	0.14**	0.40	0.78	0.59
8	499	0.05	0.17	0.00	0.21	2.57	1.73
9	737	0.01	0.59	0.18**	0.56	0.69	0.60
10	727	0.42**	0.27	0.36**	0.29	0.58	0.60
11	1412	0.06*	0.48	0.12**	0.58	1.43	0.67
12	434	0.24**	0.35	0.30**	0.39	1.07	0.32
13	1173	0.05	0.47	0.12**	0.44	1.40	0.42
14	983	0.09**	0.50	0.33**	0.44	1.00	0.49
15	2031	0.22**	0.35	0.30**	0.40	1.30	0.53
16	748	0.15**	0.43	0.41**	0.45	1.86	0.79
17	2234	0.22**	0.32	0.05	0.34	1.09	0.56
18	330	0.10	0.35	0.22**	0.43	0.89	0.55
19	1007	0.14**	0.26	0.21**	0.31	0.88	0.49
20	669	0.31**	0.29	0.46**	0.29	1.09	0.40
21	205	0.30**	0.31	0.13	0.35	0.91	0.50
22	1094	0.17**	0.48	0.25**	0.47	1.28	0.56
23	164	0.33**	0.47	0.33**	0.45	1.41	0.53
24	637	0.03	0.17	0.08	0.22	0.71	0.46

Table 5.4: Subject-independent prediction results per subject (from 5-fold CV) using the ViT-B/16 model trained on Scalp Topography images (Alpha and Theta bands, 2-second window). N: number of valid trials per subject, corr: Pearson correlation ($p<0.05$*, $p<0.01$**), MAE (s), RT (avg): subject's average actual RT (s), Dummy MAE: baseline MAE for the subject. Topo = Topography Images

Sub	N	Alpha Band (ViT Topo)		Theta Band (ViT Topo)		RT	Dummy
		corr	MAE	corr	MAE	(avg)	MAE
1	780	0.19**	0.37	0.19**	0.37	1.24	0.38
2	673	0.22**	0.27	0.22**	0.27	0.78	0.49
3	356	0.15*	0.42	0.27**	0.40	1.53	0.89
4	1356	0.42**	0.36	0.38**	0.34	1.15	0.55
5	355	0.12*	0.38	0.15*	0.36	1.13	0.43
6	617	0.33**	0.25	0.31**	0.26	1.03	0.51
7	414	0.29**	0.34	0.22**	0.33	0.78	0.59
8	499	0.13*	0.14	0.08	0.18	2.57	1.73
9	737	0.09	0.49	0.26**	0.46	0.69	0.60
10	727	0.48**	0.22	0.42**	0.24	0.58	0.60
11	1412	0.15**	0.43	0.21**	0.53	1.43	0.67
12	434	0.35**	0.26	0.41**	0.30	1.07	0.32
13	1173	0.10*	0.41	0.17**	0.38	1.40	0.42
14	983	0.28**	0.30	0.52**	0.24	1.00	0.49
15	2031	0.25**	0.32	0.33**	0.37	1.30	0.53
16	748	0.30**	0.38	0.56**	0.40	1.86	0.79
17	2234	0.26**	0.29	0.09*	0.31	1.09	0.56
18	330	0.22**	0.30	0.34**	0.38	0.89	0.55
19	1007	0.30**	0.24	0.37**	0.29	0.88	0.49
20	669	0.40**	0.27	0.55**	0.27	1.09	0.40
21	205	0.58**	0.22	0.41**	0.26	0.91	0.50
22	1094	0.20**	0.43	0.28**	0.42	1.28	0.56
23	164	0.35**	0.42	0.35**	0.40	1.41	0.53
24	637	0.08*	0.14	0.13*	0.19	0.71	0.46

with Alpha band Scalp Topographies, Subject 10 shows a correlation of 0.48** and an MAE of 0.22s with ViT-B/16, compared to 0.42** and 0.27s with ResNet18. Subject 21 achieves an outstanding correlation of 0.58** and MAE of 0.22s with ViT-B/16, a notable improvement over ResNet18 (corr=0.30**, MAE=0.31s). Similar improvements can be observed for many other subjects when comparing the Theta band results between ViT-B/16 and ResNet18. This enhanced consistency at the individual level underscores the ViT-B/16's superior ability to generalize and extract robust predictive patterns from the EEG images despite inherent inter-subject differences.

In conclusion, the experimental results robustly support the hypothesis that transforming EEG spectral features into 2D image representations and processing them with an advanced Vision Transformer (ViT-B/16) can lead to state-of-the-art performance in subject-independent driver reaction time prediction. This approach surpassed not only classical machine learning models and standard CNNs (ResNet18) applied to these images but also the specialized 1D-CNN that previously set the benchmark on 1D spectral features. Scalp Topography images emerged as a marginally more effective representation for the ViT-B/16 than PSD Matrix Images. These findings significantly advance the potential for using sophisticated vision architectures to decode complex cognitive states from EEG data.

5.4 Discussion

The empirical results presented in this chapter represent a novel contribution to the field of EEG-based cognitive state prediction by demonstrating, through rigorous subject-independent validation, the superiority of a Vision Transformer (ViT-B/16) architecture applied to 2D image representations of spectral features for predicting driver reaction time. The findings, culminating in the high performance of the ViT models, provide compelling answers to my research questions and offer significant insights into advanced strategies for decoding driver RT. This discussion will interpret these findings, focusing on why the ViT-B/16 excelled, the comparative efficacy of the image representations, the

implications for EEG-based cognitive state assessment, and the context of these results within the broader thesis narrative.

5.4.1 Vision Transformers as Superior Decoders of Image-Transformed EEG Features (RQ1)

The primary research question (RQ1) investigated whether transforming 1D EEG spectral features into 2D image representations and applying established deep learning vision architectures could lead to improved RT prediction accuracy compared to models operating directly on the original 1D features. The findings demonstrate a clear hierarchy. While a standard vision CNN (ResNet18) applied to these EEG-derived images surpassed classical machine learning models (ANN, Bayesian Ridge from Chapter 4), it did not outperform the specialized 1D-CNN. However, the introduction of the more advanced Vision Transformer (ViT-B/16) decisively shifted this balance. As shown in Table 5.2, the ViT-B/16 models, particularly when processing Scalp Topography images from Alpha and Theta band activity, achieved significantly lower Mean Absolute Errors (MAEs) and higher Pearson correlations than the 1D-CNN benchmark. For instance, the ViT-B/16 (Alpha Scalp Topo) yielded an MAE of 0.33s, an 8.3% reduction compared to the 1D-CNN's 0.36s for the same band.

This superior performance of ViT-B/16 can be attributed to its core architectural design:

- **Global Contextual Understanding via Self-Attention:** Unlike the local receptive fields of CNNs (both 1D and 2D like ResNet18), ViTs employ self-attention mechanisms across all image patches [101]. This enables the model to capture long-range dependencies and understand the global context of the input image. For EEG-derived images, this translates to an enhanced ability to identify complex, spatially distributed neural signatures (e.g., inter-regional synchronization or desynchronization patterns visible in scalp topographies) or intricate relationships across the entire channel-frequency spectrum (in PSD Matrix Images) that are indicative of the driver's cognitive state and subsequent RT. The 1D-CNN, while

effective for sequential patterns, might be less adept at modeling these holistic, non-local interactions when data is presented in a 2D format.

- **Flexible Feature Learning from Patches:** ViTs process images by dividing them into patches and learning representations from these patches and their interactions. This approach might be particularly well-suited for EEG images where salient information might not always be localized in a manner that aligns perfectly with the fixed kernels of traditional CNNs. The attention mechanism can dynamically weigh the importance of different patches (representing different scalp regions or channel-frequency blocks).
- **Effective Transfer Learning from Large-Scale Pre-training:** The ViT-B/16 models were initialized with weights pre-trained on ImageNet. The powerful and diverse feature hierarchies learned from this massive dataset provide a robust starting point for fine-tuning on more specialized domains like EEG-derived images, enabling the model to learn effectively even with a moderately sized target dataset of derived images.

The progression from classical ML, to ResNet18 on images, to ViT-B/16 on images clearly indicates that the representational power of the chosen deep learning architecture is a critical factor. The image transformation strategy for EEG data becomes maximally beneficial when paired with a vision model like ViT that can fully exploit the global structure and complex patterns within these 2D representations.

5.4.2 Optimal Image Representation for Vision Transformer Analysis (RQ2)

The second research question (RQ2) focused on which 2D EEG image representation—PSD Matrix Images or Scalp Topographies—and which vision architecture (ResNet18 or ViT-B/16) proved most effective. The results indicate that while both image types allowed the ViT-B/16 to outperform the 1D-CNN, **Scalp Topography images consistently yielded marginally superior performance** across most frequency bands and metrics

when processed by the ViT-B/16 (Table 5.2). For example, in the Alpha band, Scalp Topographies led to an MAE of 0.33s and a correlation of 0.40, compared to 0.34s and 0.38 for PSD Matrix Images.

This slight advantage for scalp topographies might be due to:

- **Neuroanatomically Relevant Spatial Structure:** Scalp topographies directly map EEG power onto a 2D representation that reflects the spatial arrangement of electrodes on the head. This explicit spatial encoding of brain activity might align well with how ViTs process images through patches, allowing the self-attention mechanism to effectively identify and relate activity patterns across different, physiologically relevant scalp regions.

- **Feature Smoothing through Band Averaging:** Scalp topographies in this study were generated from band-averaged power (power averaged across all frequency bins within a canonical band for each channel). This averaging process might act as a beneficial form of feature smoothing or noise reduction, presenting a more stable spatial signal to the ViT. In contrast, PSD Matrix Images present the power for every individual frequency bin, potentially including more fine-grained detail but also more noise or less consistently predictive bins.

It is important to note that PSD Matrix Images still enabled the ViT-B/16 to achieve excellent results, surpassing the 1D-CNN. This suggests they also contain rich, decodable information about channel-frequency interactions. The choice between these representations for future ViT-based EEG analysis might depend on the specific research question or the nature of the EEG phenomena being investigated. However, for this RT prediction task, scalp topographies appeared to be slightly more advantageous.

When comparing architectures, the ViT-B/16 was clearly superior to ResNet18 when applied to these EEG images. This reinforces that simply transforming data to an image format is not enough; the choice of a sufficiently powerful and appropriate vision model is paramount to realize the benefits of such a transformation.

5.4.3 Consistency and Variability in Subject-Level Predictions

The per-subject results for the ViT-B/16 model (Table 5.4), particularly for the Alpha and Theta band Scalp Topographies, demonstrate that the strong aggregate performance translates to improved predictions for a significant number of individual participants. When compared to the per-subject data from the ResNet18 model (Table 5.3) and the classical ANN model (Table 4.2), the ViT-B/16 generally provided lower MAEs and more consistently significant Pearson correlations. This indicates a more robust and generalizable model. For instance, many subjects who showed moderate or weak correlations with earlier models exhibited stronger and more statistically significant relationships with the ViT-B/16, suggesting it captured more reliable predictive patterns.

Nevertheless, inter-subject variability in prediction accuracy remains an inherent characteristic of EEG-based cognitive state assessment. While the ViT-B/16 model mitigates this to some extent by achieving better overall performance, some individuals are still predicted with higher accuracy than others. This persistent variability likely reflects genuine neurophysiological differences between individuals, variations in task engagement, and potentially residual uncorrected artefacts.

5.4.4 Broader Implications for EEG Analysis and Cognitive State Prediction

The findings of this chapter carry several important implications for the field of EEG analysis and the development of systems for predicting cognitive-behavioral outcomes:

1. **Viability of Advanced Vision Models for EEG Data:** This research robustly demonstrates that state-of-the-art vision architectures like Vision Transformers can be highly effective for decoding EEG data when it is appropriately transformed into an image domain. This opens up EEG analysis to a powerful class of models with proven capabilities in complex pattern recognition.

2. **Importance of Global Contextual Information:** The success of ViT-B/16 suggests that for tasks like RT prediction from pre-stimulus EEG, modeling global,

long-range dependencies in the spectro-spatial patterns of brain activity is crucial and can lead to performance gains over models focused primarily on local features.

3. **New Benchmarks and Methodological Pathways:** This study establishes a new, higher performance benchmark for pre-stimulus EEG-based RT prediction on the utilized dataset. It also validates a methodological pathway—transforming spectral EEG features into images (especially topographies) and applying ViTs—that can be explored for a wide array of other EEG-based classification and regression problems (e.g., emotion recognition, workload estimation, clinical diagnostics).

5.4.5 Limitations and Future Considerations

Despite the significant advancements demonstrated, some limitations and considerations persist:

- **Computational Demands of Vision Transformers:** ViT models, including ViT-B/16, are computationally more intensive than the 1D-CNN or ResNet18, both in terms of training resources (GPU memory, time) and inference latency. This is a critical factor for real-time applications and edge deployment (a challenge addressed for a different task in Chapter 7).

- **Data Requirements for Fine-Tuning ViTs:** While transfer learning from ImageNet is highly beneficial, ViTs generally perform best when fine-tuned on substantial target datasets. The dataset of ∼157,000 derived EEG images, while large, originates from 24 unique subjects. Larger and more diverse source EEG datasets could potentially lead to even better ViT performance.

- **Optimal EEG-to-Image Transformation:** The methods used here for generating PSD Matrix Images and Scalp Topographies are standard but represent just two possibilities. Further research into optimizing EEG-to-image conversion techniques specifically for ViT processing could yield additional performance gains.

Future work could explore more computationally efficient ViT variants (e.g., MobileViT,

as explored in Chapter 7 for a different task, or other compact transformer designs) applied to these EEG images. Investigating different patch sizes, image resolutions, and longer fine-tuning schedules might also refine performance. Furthermore, exploring the application of ViTs to other types of EEG-derived images (e.g., from connectivity measures or time-frequency representations that retain more temporal detail within the image itself) is a promising direction.

5.5 Conclusion

This chapter has definitively established that transforming pre-stimulus EEG spectral features into 2D image representations and subsequently processing them with an advanced Vision Transformer (ViT-B/16) architecture can achieve a new state-of-the-art in subject-independent driver reaction time prediction accuracy. The ViT-B/16 models, when applied to either PSD Matrix Images or, marginally more effectively, Scalp Topographies, significantly outperformed not only classical machine learning techniques and standard vision CNNs (ResNet18) operating on these images, but also the specialized 1D-CNN that had previously set the performance benchmark on 1D spectral feature vectors (RQ1).

Specifically, the ViT-B/16 model processing Alpha band Scalp Topography images yielded a mean MAE of 0.33s and a mean Pearson correlation of 0.40, representing a substantial improvement in predictive capability. While Scalp Topographies demonstrated a slight, consistent advantage over PSD Matrix Images as inputs for the ViT-B/16 (RQ2), both image types proved to be effective mediums for conveying RT-predictive information to the transformer architecture.

The success of this approach underscores the remarkable capacity of Vision Transformers to model global context and learn complex, discriminative patterns from appropriately structured data, even when that data originates from a non-visual modality like EEG. These findings strongly advocate for the continued exploration of image-based transformations coupled with state-of-the-art vision architectures as a powerful paradigm for

advancing EEG-based cognitive state assessment and brain-computer interface technologies. The performance levels achieved here offer a robust foundation for future research aiming to further refine these techniques for applications in driver safety and beyond.

Chapter 6

Multimodal Transformer-Based Fusion of EEG and Vision for Driver Drowsiness Detection

6.1 Introduction

The preceding chapters of this thesis (Chapters 4 and 5) have rigorously investigated the potential of Electroencephalography (EEG) signals to predict driver reaction time, a critical correlate of vigilance and cognitive preparedness. These studies established that pre-stimulus EEG features, particularly from specific frequency bands and time windows, can indeed offer predictive power. However, driver safety is also profoundly impacted by more holistic states like drowsiness, which manifest through a complex interplay of internal physiological changes and external behavioural cues. Detecting drowsiness directly and reliably is paramount for preventing accidents, as highlighted by alarming statistics linking drowsy driving to a significant number of road fatalities [25, 23, 6].

While unimodal approaches have been extensively explored for drowsiness detection, each carries inherent limitations. EEG-based methods, though providing direct access to brain activity [13, 106, 42, 43], can be susceptible to artefacts, exhibit high inter-subject variability, and may be perceived as intrusive for continuous in-vehicle monitoring

[63, 64]. Vision-based systems, which analyze facial features such as eye closure, blink patterns, and yawning frequency [17, 38, 39, 40], offer a non-intrusive alternative but can be compromised by variable lighting conditions, occlusions, and may only capture overt behavioural manifestations rather than the subtle onset of cognitive impairment [16].

Recognizing these individual shortcomings, recent research has increasingly turned towards multimodal approaches, hypothesizing that the fusion of complementary data streams can lead to more robust and accurate drowsiness detection systems [18, 19, 105]. The internal neurophysiological state captured by EEG and the external behavioural cues monitored by vision systems are prime candidates for such fusion. The challenge, however, lies in effectively integrating these diverse signals to exploit their synergistic potential.

This chapter addresses this challenge by exploring advanced multimodal fusion techniques, with a particular focus on transformer-based architectures. Transformers, originally developed for natural language processing [110], have demonstrated remarkable success in modeling long-range dependencies and capturing complex feature interactions, and have recently been adapted for computer vision (Vision Transformers, ViTs [111]) and multimodal tasks [112]. Their capacity for self-attention and cross-modal attention makes them well-suited for learning intricate relationships between different data modalities like EEG and video.

The primary goal of this chapter is to investigate the efficacy of fusing synchronized EEG and facial video data for subject-independent driver drowsiness classification. I systematically evaluate different fusion strategies, progressing from simple feature concatenation to sophisticated transformer-based models, culminating in an end-to-end architecture that processes raw EEG and video data jointly. This investigation is guided by the following research questions:

1. **RQ1: How effective are standard unimodal deep learning models (EEGNet for EEG, ResNet18/Vision Transformer for Vision) at classifying driver drowsiness into 'Alert' versus 'Drowsy' states in a subject-independent setting using the Tobii multimodal dataset?** This establishes crucial perfor-

mance baselines for each modality.

2. **RQ2: Can simple feature-level fusion of EEG and vision features, using a classical classifier like Bayesian Ridge Classification, improve drowsiness classification accuracy compared to the best performing unimodal baseline?** This tests the benefit of a basic fusion approach.

3. **RQ3: Can transformer-based multimodal fusion models, particularly an end-to-end architecture processing raw EEG and vision data, significantly outperform both unimodal models and simpler fusion strategies, thereby demonstrating superior synergistic integration of the EEG and vision modalities?** This assesses the potential of advanced deep learning fusion.

To address these questions, I have utilized the comprehensive Tobii multimodal driving dataset (introduced in Chapter 3, Section 3.3). This dataset is specifically designed for drowsiness research, featuring data from 79 participants recorded during distinct 'Alert' (10 AM) and 'Drowsy' (3 AM) driving simulation sessions, with subjective drowsiness levels confirmed by Karolinska Sleepiness Scale (KSS) assessments. 1-second data segments are classified as either 'Alert' or 'Drowsy'.

This chapter will first detail the methodologies for data preprocessing, the unimodal baseline models, and the various fusion architectures. Subsequently, it will present a comparative analysis of their performance, aiming to identify the most effective strategy for multimodal drowsiness detection. The findings are expected to contribute to the development of more reliable driver monitoring systems by elucidating the benefits of advanced deep learning fusion techniques.

6.2 Methods

The methodological framework detailed in this section outlines the experimental procedures undertaken to investigate multimodal fusion for driver drowsiness detection. This includes a description of the Tobii dataset, the specific preprocessing pipelines for both

Electroencephalogram (EEG) and vision data, the architectures of the unimodal baseline models, and the design of the various feature-level and end-to-end multimodal fusion strategies, with a particular emphasis on the proposed transformer-based architectures.

6.2.1 Dataset and Experimental Design (Tobii Dataset)

This research leverages the Tobii multimodal driving fatigue dataset, comprehensively described in Chapter 3, Section 3.3. Key aspects pertinent to the experiments in this chapter are reiterated here:

- **Participants:** The study utilized data from 79 participants. These participants were selected from an initial pool of 100, with 21 being excluded due to issues such as excessive noise, artefacts in EEG recordings (e.g., significant facial muscle interference), or problems related to electrode connections which compromised signal integrity. The participant pool exhibited an age range of 18 to 71 years, with a gender distribution of approximately 30% women and 70% men.

- **Experimental Conditions and Fatigue Induction:** Each participant engaged in a driving simulation task during two distinct sessions designed to capture varying states of alertness:

    1. An 'Alert' state session, conducted at 10 AM, after participants had a normal night's sleep.

    2. A 'Drowsy' state session, conducted at 3 AM, after participants had remained awake for approximately 24 hours under supervision within the research facility. This protocol was designed to induce significant levels of drowsiness due to circadian factors and sleep deprivation.

- **Data Modalities Utilized:** For the investigations in this chapter, synchronized EEG data and RGB facial video data were employed.

- **Ground Truth Labeling for Drowsiness Classification:** The core task is a binary classification problem. Each 1-second segment of recorded data was labeled

as either 'Alert' (originating from the 10 AM session) or 'Drowsy' (originating from the 3 AM session). The Karolinska Sleepiness Scale (KSS) [12] was administered, with a board-certified neurologist providing scores for each one-second interval. A statistical comparison (independent samples t-test) of these KSS scores revealed a significantly higher mean KSS rating at 3 AM (mean KSS = 9.08 ± 0.99) compared to 10 AM (mean KSS = 6.38 ± 1.61), with t(df) = 889.56, $p < 0.00001$. This statistically robust difference validates the use of session time as a proxy for 'Alert' versus 'Drowsy' states. The distribution of KSS scores for these sessions is illustrated in Figure 6.1.

- **Ethical Considerations and Data Handling:** Written informed consent was obtained from all participants. Data management protocols ensured participant confidentiality through anonymization (where feasible) and secure storage on isolated servers with managed access.

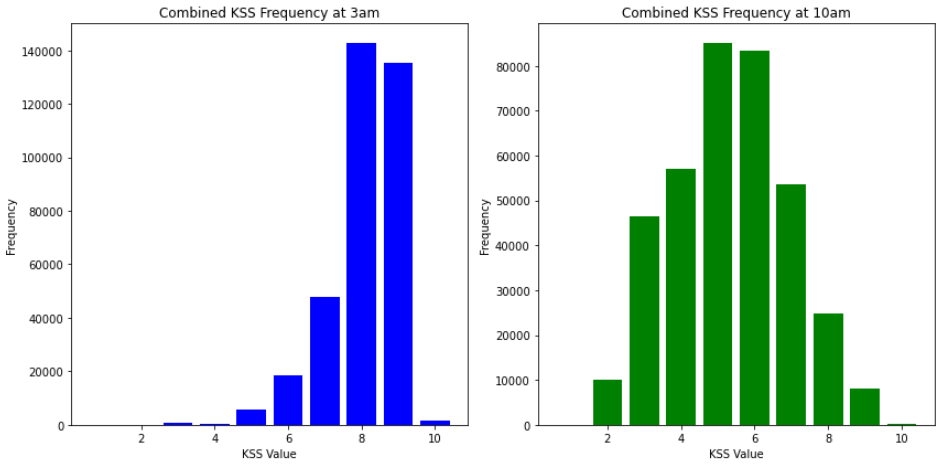


Figure 6.1: Distribution of Karolinska Sleepiness Scale (KSS) scores from the Tobii dataset, illustrating distinct distributions for the 10 AM ('Alert', depicted in the right panel of the original source, showing lower scores) and 3 AM ('Drowsy', depicted in the left panel of the original source, showing higher scores) sessions. The 3 AM sessions clearly exhibit significantly higher KSS scores.

6.2.2 Data Preprocessing Pipelines

Preprocessing pipelines were applied to the EEG and vision data streams to prepare them for input into the deep learning models.

EEG Data Preprocessing

The 32-channel EEG signals, acquired at a sampling rate of 500 Hz, were subjected to the following sequence of preprocessing operations:

1. **Band-Pass Filtering:** The continuous EEG recordings were filtered into four canonical frequency bands. This was achieved using a zero-phase Finite Impulse Response (FIR) filter, designed with a Hamming window, as implemented in MNE-Python's `filter_data` function [56]. The specific bands isolated were:
    - Delta (δ): 1 Hz to 4 Hz
    - Theta (θ): 4 Hz to 8 Hz
    - Alpha (α): 8 Hz to 12 Hz
    - Beta (β): 13 Hz to 30 Hz
2. **Segmentation:** Following artefact correction, the continuous EEG data for each band was segmented into non-overlapping 1-second windows. This temporal resolution aligns with the 1-second KSS annotations and is suitable for capturing relatively rapid state fluctuations.
3. **Normalization:** The amplitude values within each 1-second EEG segment were normalized. This typically involves standardization (e.g., Z-score normalization per channel or across channels for the segment) to ensure that the input data has a consistent scale, which is beneficial for training neural networks.

Vision Data Preprocessing

The RGB facial video data, captured at 30 frames per second, was processed to align with the EEG data and meet the input requirements of the vision models:

1. **Frame Extraction and Synchronization:** For each 1-second EEG segment, a corresponding set of video frames was extracted. Given the 30 FPS recording rate, each 1-second segment corresponds to 30 frames.

2. **Resizing:** The extracted facial image frames were uniformly resized to 224×224 pixels. This is a standard input dimension for many pre-trained CNNs and Vision Transformers, including ResNet18 and ViT-Base.

3. **Normalization:** Pixel intensity values of the resized frames were normalized. Common practice involves standardizing them using the mean and standard deviation of the dataset on which the vision models were pre-trained (e.g., ImageNet statistics).

The dataset resulting from this preprocessing comprised 284,400 one-second instances, each consisting of a preprocessed EEG segment (for each band) and a corresponding preprocessed facial video frame (or frame sequence). The dataset was balanced, with 142,200 instances for the 'Alert' class and 142,200 instances for the 'Drowsy' class.

6.2.3 Unimodal Baseline Models (Addressing RQ1)

To establish performance benchmarks for each modality independently, specific deep learning architectures were selected and trained.

EEG Baseline Architecture: EEGNet

For classifying drowsiness based solely on EEG data, the EEGNet architecture, proposed by Lawhern et al. [60], was employed. EEGNet is a compact Convolutional Neural Network specifically engineered for EEG-based classification tasks. It features a sequence of convolutional blocks, including depthwise and separable convolutions, designed to efficiently learn both spatial (across channels) and temporal (within the 1-second window) features from EEG signals. Independent EEGNet models were trained and evaluated for each of the four preprocessed frequency bands (Delta, Theta, Alpha, and Beta). The input to each model was the 1-second multi-channel EEG data corresponding to its designated band.

Vision Baseline Architectures: ResNet18 and Vision Transformer (ViT)

Two distinct deep learning architectures were utilized as baselines for vision-based drowsiness classification:

1. **ResNet18:** A ResNet18 model [97], pre-trained on the ImageNet dataset, was fine-tuned for the binary drowsiness classification task. The original classification layer was replaced with a new one suited for two classes (Alert/Drowsy). ResNet18 serves as a robust and widely adopted CNN baseline for image classification.

2. **Vision Transformer (ViT):** A pre-trained Vision Transformer (ViT-Base model) [111], also initialized with ImageNet weights, was fine-tuned. ViTs operate by dividing an image into a sequence of fixed-size patches, linearly embedding them, and then processing this sequence with a standard transformer encoder. This architecture allows for modeling global relationships between image regions via self-attention mechanisms.

Both vision models received the 224×224 normalized facial image frames as input and were trained to perform binary classification.

6.2.4 Feature-Level Multimodal Fusion Strategies

Two strategies for feature-level fusion were investigated, where features are first extracted independently from each modality and then combined for a final classification decision.

Feature Extraction for Fusion Pipelines

High-level feature representations were extracted from the penultimate layers of the trained unimodal baseline models:

- **EEG Features:** Feature vectors were obtained from the trained EEGNet models (one set of features per frequency band).

- **Vision Features:** Feature vectors were obtained from the trained ResNet18 model. ResNet18 was used as the feature extractor for these initial feature-level fusion

experiments, aligning with common practice for establishing simpler fusion baselines before moving to more complex end-to-end systems with ViT.

Simple Concatenation with Bayesian Ridge Classification (Addressing RQ2)

The first feature-level fusion approach involved a direct concatenation of the extracted EEG features (for a given band) and the ResNet18 vision features into a single, combined feature vector. This fused vector was then used as input to a Bayesian Ridge Classification model [109]. This method was selected as a representative simple linear fusion technique. Hyperparameters of the Bayesian Ridge classifier (‘alpha_1‘, ‘alpha_2‘, ‘lambda_1‘, ‘lambda_2‘) were carefully tuned via a grid search with 5-fold cross-validation on the training data splits to optimize its performance, with the optimal set being ‘alpha_1 = 1e-6‘, ‘alpha_2 = 1e-6‘, ‘lambda_1 = 1e-6‘, and ‘lambda_2 = 1e-6‘.

Transformer-Based Feature-Level Fusion (Addressing RQ3, Part 1)

To explore a more sophisticated mechanism for integrating the pre-extracted features, a transformer-based fusion model was developed. This model, conceptually illustrated in Figure 6.2, processes the feature vectors obtained from EEGNet (per band) and ResNet18. The architecture includes:

- **Modality-Specific Linear Projections:** Initial linear layers to project the EEG and vision feature vectors into a common embedding dimension.
- **Positional Encodings:** Added to embeddings to maintain positions.
- **Cross-Modal Attention Layers:** A stack of transformer encoder layers employing cross-modal attention. This allows, for instance, the vision feature representation to be conditioned on the EEG feature representation and vice-versa, enabling the model to learn weighted importance and interactions between features from the different modalities.
- **Fusion and Classification Head:** The resulting attended multimodal representations are then typically fused (e.g., by concatenation or pooling) and passed through

an MLP classification head to produce the final drowsiness prediction.

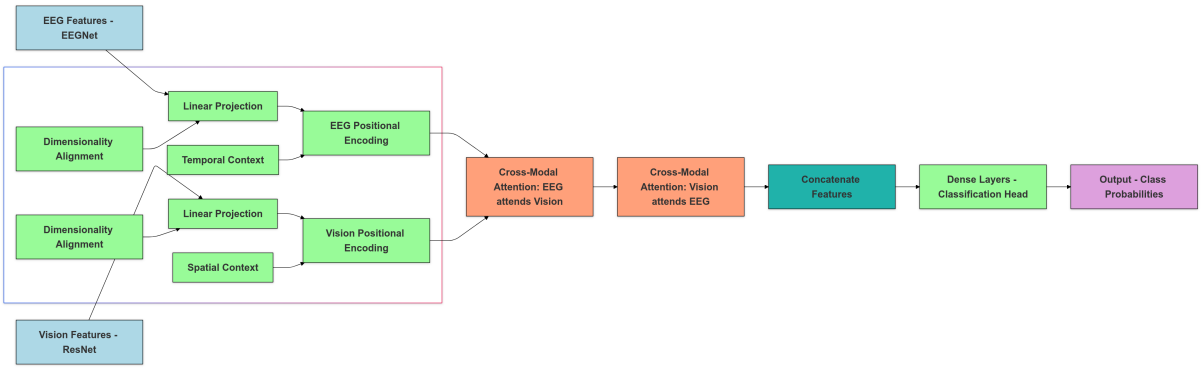


Figure 6.2: Conceptual diagram of the transformer-based feature-level fusion model. Features extracted by EEGNet (for EEG) and ResNet18 (for vision) are fed into separate initial encoders (linear projections). These are then processed by cross-modal attention transformer layers, followed by a fusion mechanism and a final classification head.

6.2.5 End-to-End Multimodal Transformer (Addressing RQ3, Part 2)

The most advanced fusion strategy investigated was an end-to-end multimodal transformer model. This architecture is designed to learn feature representations from the raw EEG and vision data streams simultaneously and to optimize their fusion jointly within a single, unified network. The conceptual design is shown in Figure 6.3. Key components include:

- **EEG Transformer Encoder:** This module directly processes the 1-second raw multi-channel EEG time-series data (per frequency band). It consists of multiple transformer encoder blocks that use self-attention mechanisms to model temporal dependencies and inter-channel relationships within the EEG segment, learning a rich representation without reliance on handcrafted features or prior CNN-based feature extraction like EEGNet.
- **Vision Transformer (ViT) Encoder:** This module processes the raw 224×224 facial video frames. It is based on a pre-trained ViT-Base architecture [111], whose weights are fine-tuned during the end-to-end training. The ViT encoder captures global spatial context from the input image patches.

- **Cross-Modal Attention Layers:** After the modality-specific encoders generate representations for EEG and vision, these representations are fed into a series of cross-modal attention layers. These layers are the core of the fusion process, allowing bidirectional conditioning where EEG features can attend to vision features and vice-versa. This enables the model to learn complex, synergistic relationships and identify complementary cues across the two modalities.

- **Classification Head:** The final, fused multimodal representation, enriched by cross-modal interactions, is passed to an MLP-based classification head which outputs the probability for the 'Drowsy' class.

This end-to-end learning paradigm offers the potential for discovering optimal feature hierarchies and fusion strategies that are specifically tailored to the drowsiness detection task, potentially overcoming limitations of pipelined approaches where feature extraction and fusion are separated.

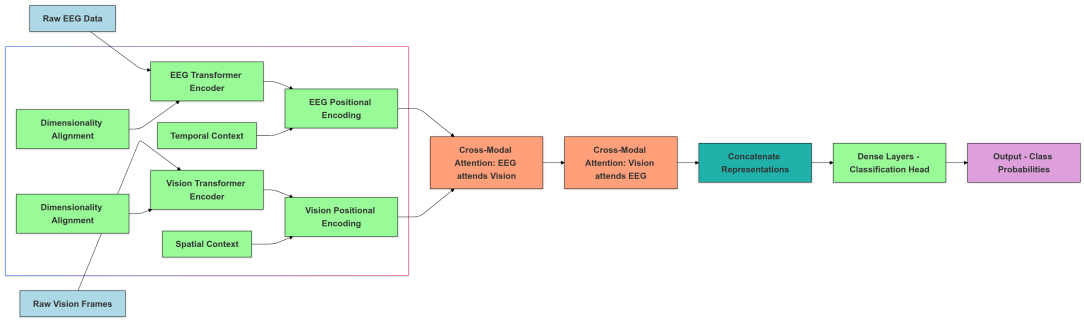


Figure 6.3: Conceptual diagram of the end-to-end multimodal transformer model. Raw EEG data is processed by a dedicated EEG Transformer Encoder, while raw vision frames are processed by a Vision Transformer (ViT) Encoder. The outputs from these modality-specific encoders are then integrated using cross-modal attention layers before being passed to a final classification head.

6.2.6 Training, Evaluation, and Implementation Details

A consistent and rigorous protocol was maintained for training and evaluating all models:

- **Subject-Independent Cross-Validation:** All models were trained and evaluated using a **5-fold cross-validation** scheme. The 79 participants were partitioned into

5 distinct folds. In each iteration, data from participants in one fold constituted the test set, while data from participants in the remaining four folds formed the training set. This strict subject separation is critical for assessing the model's ability to generalize to previously unseen individuals.

- **Training Parameters for Deep Learning Models:**
  - **Optimizer:** The Adam optimizer [133] was consistently used.
  - **Learning Rate:** An initial learning rate of 1×10^{-4} was set for training.
  - **Loss Function:** Binary Cross-Entropy loss was employed, suitable for the binary (Alert/Drowsy) classification task.
  - **Batch Size:** A mini-batch size of 32 samples was used.
  - **Epochs and Early Stopping:** Models were trained for a maximum of 10 epochs. An early stopping criterion was implemented, monitoring performance (e.g., validation loss or accuracy) on a held-out validation subset of the training folds. Training ceased if no improvement was observed for a set number of consecutive epochs, and the model weights from the best-performing epoch were retained.
  - **Initialization and Fine-Tuning:** Vision model components (ResNet18, ViT-Base) were initialized with weights pre-trained on ImageNet [132] and subsequently fine-tuned. For the end-to-end transformer, the last 4 transformer blocks of the ViT vision encoder were fine-tuned. The EEG Transformer Encoder weights were initialized using Xavier initialization [134].
- **Software Implementation:** All deep learning models were implemented using the PyTorch deep learning framework [135]. Training was accelerated using NVIDIA GPUs.
- **Performance Metrics:** Model efficacy was assessed using standard binary classification metrics, averaged across the 5 cross-validation folds:
  - Accuracy

- Precision (for the 'Drowsy' class)
- Recall (Sensitivity, for the 'Drowsy' class)
- Area Under the Receiver Operating Characteristic Curve (AUC-ROC)

This comprehensive methodological approach ensures that the comparisons between different unimodal and multimodal strategies are robust and that the reported results accurately reflect the models' generalization capabilities.

6.3 Results

This section presents the empirical findings from the application of the various unimodal and multimodal modeling strategies to the Tobii dataset for driver drowsiness detection. The performance is reported based on the 5-fold subject-independent cross-validation protocol, with results averaged across folds. I first detail the performance of the unimodal baseline models (RQ1), followed by an evaluation of the feature-level fusion techniques (RQ2 and part of RQ3), and conclude with the results from the end-to-end multimodal transformer architecture (fully addressing RQ3).

6.3.1 Performance of Unimodal Baseline Models (RQ1)

To establish individual modality performance benchmarks, EEGNet models were trained and tested for each EEG frequency band, and both ResNet18 and Vision Transformer (ViT-Base) models were trained and tested for the vision modality. The classification results, including Accuracy, Precision, Recall, and AUC-ROC, are presented in Table 6.1.

The unimodal EEGNet models exhibited limited classification accuracy. The Theta and Alpha bands yielded the highest EEG-only accuracies, at 58.11% and 58.21% respectively. These figures, while marginally better than chance, underscore the significant challenge of robustly classifying drowsiness from 1-second EEG segments in a subject-independent manner. The delta band's performance was lower, and the beta band's

Table 6.1: Subject-independent drowsiness classification performance (mean across 5 folds) for unimodal EEGNet models (per frequency band) and vision models (ResNet18, ViT-Base) on the Tobii dataset.

Model	Accuracy (%)	Precision	Recall	AUC-ROC
EEG Modality (EEGNet)				
EEG (Delta Band)	53.27	0.528	0.541	0.5428
EEG (Theta Band)	58.11	0.574	0.592	0.6268
EEG (Alpha Band)	58.21	0.576	0.592	0.6183
EEG (Beta Band)	51.50	0.520	0.508	0.4867
Vision Modality				
Vision (ResNet18)	83.02	0.810	0.855	0.9151
Vision (ViT-Base)	**85.78**	**0.801**	**0.962**	**0.9250**

accuracy was close to chance level, with an AUC-ROC below 0.5, suggesting it provided little discriminative information in this context.

In contrast, the vision-based models demonstrated substantially greater predictive power. The ResNet18 model achieved an accuracy of 83.02% and an AUC-ROC of 0.9151. The Vision Transformer (ViT-Base) model further improved upon these results, attaining an accuracy of 85.78% and an AUC-ROC of 0.9250. The ViT-Base also exhibited a very high recall of 0.962, indicating its proficiency in identifying true drowsy instances. These results clearly establish that, within a unimodal framework, visual cues from facial video are considerably more discriminative for subject-independent drowsiness detection than EEG spectral features from short segments. The ViT-Base model served as the most performant unimodal baseline.

6.3.2 Performance of Feature-Level Fusion Strategies (RQ2)

Next, I evaluated whether combining features extracted from the unimodal EEGNet and ResNet18 models could lead to improved drowsiness classification.

Simple Fusion with Bayesian Ridge Classification

The initial feature-level fusion involved concatenating EEGNet features (from each band separately) with ResNet18 vision features, and then classifying these combined vectors using Bayesian Ridge Classification. The performance metrics for this approach are shown

in Table 6.2.

Table 6.2: Subject-independent drowsiness classification performance (mean across 5 folds) using Bayesian Ridge Classification on concatenated EEGNet (per band) and ResNet18 vision features.

(EEG Band + Vision)	Accuracy (%)	Precision	Recall	AUC-ROC
Delta Band + Vision (RS18)	80.46	0.792	0.844	0.8869
Theta Band + Vision (RS18)	80.47	0.794	0.841	0.8877
Alpha Band + Vision (RS18)	80.49	0.796	0.838	0.8879
Beta Band + Vision (RS18)	80.85	0.785	0.868	0.8940

The results from the Bayesian Ridge fusion strategy indicate that this simple concatenation approach yielded accuracies ranging from 80.46% to 80.85%. While these scores are substantially better than the EEG-only unimodal performances, they are notably lower than the accuracy achieved by the ResNet18 vision-only model (83.02%) and significantly lower than the ViT-Base vision-only model (85.78%). This finding suggests that straightforward linear fusion of pre-extracted features from a weaker modality (EEGNet features) with those from a stronger one (ResNet18 vision features) did not result in a synergistic improvement; rather, it appeared to dilute the predictive power of the stronger vision modality.

Transformer-Based Feature-Level Fusion

A more advanced feature-level fusion was implemented using a transformer architecture incorporating cross-modal attention mechanisms, operating on the same sets of extracted EEGNet (per band) and ResNet18 vision features. The performance of this model is presented in Table 6.3.

Table 6.3: Subject-independent drowsiness classification performance (mean across 5 folds) using a transformer model with cross-modal attention on extracted EEGNet (per band) and ResNet18 vision features.

(EEG Band + Vision)	Accuracy (%)	Precision	Recall	AUC-ROC
Delta Band + Vision (TF features)	80.87	0.769	0.881	0.8848
Theta Band + Vision (TF features)	83.33	0.802	0.884	0.8995
Alpha Band + Vision (TF features)	83.10	0.788	0.905	0.9132
Beta Band + Vision (TF features)	81.71	0.770	0.904	0.9084

The transformer-based feature fusion model demonstrated better integration capabilities than the simple Bayesian Ridge approach. Accuracies improved, ranging from 80.87% to 83.33%. The fusion involving Theta band EEG features with ResNet18 vision features achieved an accuracy of 83.33% and an AUC-ROC of 0.8995. Similarly, the Alpha band EEG fusion yielded an accuracy of 83.10% and an AUC-ROC of 0.9132. These results are competitive with, and in the case of Theta band fusion, slightly exceed, the performance of the ResNet18 vision-only baseline (83.02% accuracy, 0.9151 AUC-ROC). This indicates that the transformer's attention mechanisms were more adept at identifying and leveraging useful interactions between the pre-extracted EEG and vision features. However, even this more sophisticated feature-level fusion did not manage to outperform the strongest unimodal baseline, the ViT-Base vision-only model (85.78% accuracy, 0.9250 AUC-ROC). This reinforces the notion that the full potential of multimodal fusion might be constrained when operating on features extracted from independently optimized unimodal networks.

6.3.3 Performance of End-to-End Multimodal Transformer (RQ3)

The pinnacle of the fusion strategies explored was the end-to-end multimodal transformer. This architecture was designed to learn feature representations from raw EEG time-series segments and raw facial video frames concurrently and to optimize their fusion through integrated cross-modal attention layers. The EEG data (per band) was processed by a dedicated EEG Transformer Encoder, while the vision data was handled by a fine-tuned ViT-Base encoder. Table 6.4 details the classification performance achieved by this end-to-end model.

Table 6.4: Subject-independent drowsiness classification performance (mean across 5 folds) using the end-to-end multimodal transformer, fusing raw EEG (per frequency band) with raw vision data (processed by a ViT-Base encoder).

(EEG Band + Vision)	**Accuracy (%)**	**Precision**	**Recall**	**AUC-ROC**
Alpha Band + Vision (End-to-End TF)	88.48	0.861	0.925	0.9265
Theta Band + Vision (End-to-End TF)	**91.00**	**0.883**	**0.951**	**0.9634**
Beta Band + Vision (End-to-End TF)	76.43	0.724	0.875	0.8314
Delta Band + Vision (End-to-End TF)	88.38	0.847	0.945	0.9266

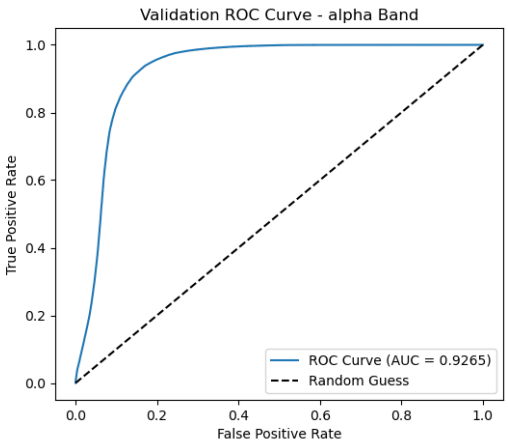
The end-to-end multimodal transformer architecture yielded a substantial improvement in drowsiness classification performance, significantly surpassing all previous approaches. The fusion of Theta band EEG data with vision data achieved the highest accuracy of **91.00%** and an AUC-ROC of **0.9634**. This represents a marked increase over the best unimodal vision model (ViT-Base: 85.78% accuracy, 0.9250 AUC-ROC) and a considerable gain over the best feature-level fusion result (Theta band EEG + Vision with Transformer: 83.33% accuracy, 0.8995 AUC-ROC).

The fusion involving Alpha band EEG data also demonstrated strong performance, reaching an accuracy of 88.48% and an AUC-ROC of 0.9265, also outperforming the ViT-Base vision-only model. The Delta band EEG fusion performed similarly to the Alpha band fusion. Consistent with earlier findings, the fusion incorporating Beta band EEG data yielded the lowest accuracy (76.43%) among the end-to-end models, though it was still notably better than its unimodal EEG counterpart.

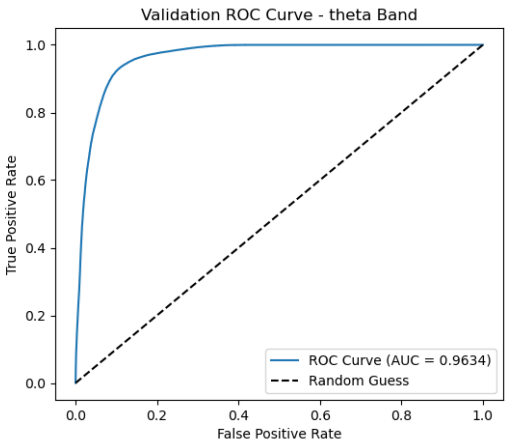
These results compellingly demonstrate the superiority of the end-to-end learning paradigm for multimodal fusion. By allowing the model to jointly learn optimal feature representations from each modality (raw EEG and raw vision) and simultaneously learn how to best integrate these representations via its cross-modal attention mechanisms, the end-to-end transformer unlocked synergistic benefits that were not fully realized by the feature-level fusion methods. The significant performance uplift, particularly with Theta and Alpha band EEG, confirms that EEG data, when appropriately fused with strong visual cues in an end-to-end fashion, provides valuable complementary information for robust driver drowsiness detection.

The excellent discriminative capability of the end-to-end models is further visualized by their Receiver Operating Characteristic (ROC) curves, presented in Figure 6.4. The curves for the Alpha and Theta band fusions, in particular, are positioned very close to the top-left corner of the ROC space, indicating high true positive rates with low false positive rates across various decision thresholds.

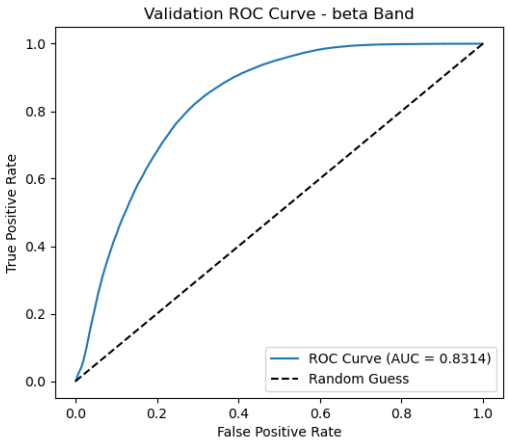
In conclusion, the empirical results clearly map a progression of performance. Unimodal EEG models show limited utility for subject-independent drowsiness classification



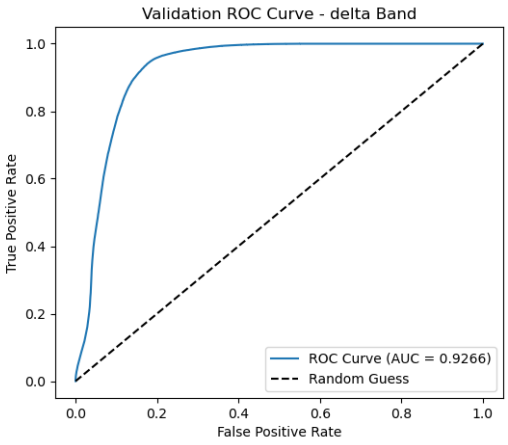
(a) Alpha Band EEG + Vision (End-to-End TF)



(b) Theta Band EEG + Vision (End-to-End TF)



(c) Beta Band EEG + Vision (End-to-End TF)



(d) Delta Band EEG + Vision (End-to-End TF)

Figure 6.4: AUC-ROC curves (averaged over 5 folds) for the end-to-end multimodal transformer, illustrating drowsiness classification performance when fusing vision data with EEG from the Alpha, Theta, Beta, and Delta frequency bands respectively. The fusion incorporating Theta band EEG with vision (b) exhibits the highest AUC value, signifying superior discrimination between Alert and Drowsy states.

from short segments. Unimodal vision models, especially ViT-Base, offer strong standalone performance. Feature-level fusion methods provide some benefits over unimodal EEG but generally fail to significantly surpass robust vision-only baselines. It is the end-to-end multimodal transformer architecture that most effectively capitalizes on the complementary nature of EEG (particularly Theta and Alpha bands) and vision data, leading to state-of-the-art drowsiness detection accuracy.

6.4 Discussion

The experimental results presented in this chapter provide significant insights into the efficacy of various unimodal and multimodal strategies for subject-independent driver drowsiness detection. The progression from simple unimodal baselines to sophisticated end-to-end multimodal transformers reveals a clear hierarchy of performance and underscores the potential of advanced deep learning techniques for integrating diverse data streams.

6.4.1 Interpreting Unimodal Performance and Modality Strengths (RQ1)

The initial unimodal baseline evaluations (Table 6.1) clearly established that vision-based cues, as processed by deep learning models like ResNet18 and particularly Vision Transformer (ViT-Base), are substantially more discriminative for drowsiness detection than EEG spectral features from short (1-second) segments in a subject-independent context. The accuracies achieved by EEGNet models (maxing out at $\approx$58%) were only modestly above chance level. This relatively low performance for EEG alone in such a challenging subject-independent, short-epoch scenario aligns with known difficulties in EEG-based classification, including high inter-subject variability in signal characteristics, susceptibility to noise, and the often subtle expression of drowsiness in brief EEG epochs without longer temporal context [136, 63]. While Theta and Alpha band activity are well-established correlates of alertness [14, 15], their utility for direct classification

from isolated 1-second segments appears limited without more advanced modeling or contextual information.

Conversely, the ViT-Base vision model achieved an accuracy of 85.78% and an AUC-ROC of 0.9250. This robust performance highlights the rich information contained in facial visual cues—such as PERCLOS (percentage of eye closure), blink rate and duration, yawning frequency, head pose, and subtle changes in facial muscle tone—that are overtly indicative of drowsiness [16, 137]. Modern vision architectures like ViT are highly effective at learning these complex visual patterns directly from image data. This disparity underscores that, as standalone modalities for this task, vision provides a much stronger predictive signal than short-segment EEG.

6.4.2 Limitations of Feature-Level Fusion Approaches (RQ2 and RQ3 Part 1)

The investigation into feature-level fusion strategies yielded mixed but ultimately revealing results. The simple concatenation of EEGNet-derived features with ResNet18-derived vision features, followed by a Bayesian Ridge Classifier (Table 6.2), failed to improve upon the ResNet18 vision-only baseline and performed considerably worse than the ViT-Base vision-only model. This outcome suggests that naive, linear fusion of features from modalities with such disparate individual predictive power may not be effective; the weaker EEG features might even introduce noise or fail to add significant complementary information that a linear classifier can exploit when combined with strong vision features.

The transformer-based feature-level fusion model (Table 6.3), which employed cross-modal attention mechanisms on the same pre-extracted EEGNet and ResNet18 features, demonstrated improved integration capabilities over the Bayesian Ridge approach. Accuracies increased, and the fusion involving Theta or Alpha band EEG features with vision reached performance levels (e.g., 83.33% accuracy for Theta fusion) comparable to the ResNet18 vision-only baseline. This indicates that the attention mechanism was somewhat successful in identifying and weighting relevant cross-modal feature interactions. However, this approach still did not surpass the strongest unimodal baseline (ViT-

Base). This limitation likely arises because the features themselves were extracted by unimodal networks (EEGNet, ResNet18) optimized for their respective individual tasks, not specifically for later fusion. Information critical for synergistic fusion might be lost or sub-optimally represented in these pre-extracted features, constraining the potential of even a sophisticated fusion module like a transformer. This addresses RQ2 by showing simple fusion is insufficient and partially addresses RQ3 by indicating that feature-level transformer fusion, while better, is also constrained.

6.4.3 The Efficacy and Significance of End-to-End Multimodal Transformer Fusion (RQ3)

The most significant finding of this chapter is the superior performance of the end-to-end multimodal transformer architecture (Table 6.4). By processing raw EEG time-series and raw vision frames directly and learning feature representations and their fusion jointly, this model achieved a classification accuracy of up to 91.00% and an AUC-ROC of 0.9634 (when fusing Theta band EEG with vision). This represents a substantial improvement over all other methods, including the powerful ViT-Base vision-only baseline (85.78% accuracy).

This success robustly answers RQ3, demonstrating that a carefully designed end-to-end multimodal transformer can indeed significantly outperform both unimodal models and simpler fusion strategies, effectively leveraging the synergy between EEG and vision. Several factors likely contribute to this:

- **Joint Optimization:** The end-to-end model learns feature extractors for both EEG (via its EEG Transformer Encoder) and vision (via the fine-tuned ViT Encoder) that are optimized not just for unimodal discrimination but specifically for their utility in the context of the other modality and the final fusion task. This avoids the potential information bottleneck of using pre-extracted features from independently trained networks.

- **Powerful Modality-Specific Encoders:** The use of dedicated transformer en-

coders for both EEG time-series and vision frames allows each modality to be processed by an architecture well-suited to its data structure. The EEG Transformer can capture long-range temporal dependencies within the 1-second EEG segment, while the ViT captures global spatial context in the image.

- **Effective Cross-Modal Attention:** The integrated cross-modal attention layers enable deep, bidirectional interactions between the learned EEG and vision representations. This allows the model to, for example, weigh visual features differently based on concurrent EEG patterns, or vice-versa, capturing nuanced cross-modal correlations indicative of drowsiness that simpler fusion methods might miss.

- **Complementarity of Information:** The results strongly suggest that EEG and vision provide complementary information regarding drowsiness. While vision captures overt behavioral signs, EEG (particularly Theta and Alpha bands) reflects underlying neurophysiological shifts in brain state associated with alertness and fatigue [48, 130]. The end-to-end model appears highly effective at integrating these internal and external indicators. The strong performance with Theta band fusion, for instance, aligns with the known increase in theta activity during drowsiness [15].

The performance achieved (91.00% accuracy) is highly competitive and represents a significant step towards robust, automated drowsiness detection. This level of accuracy, achieved in a rigorous subject-independent validation, is particularly promising for real-world applications.

6.4.4 Comparison with Prior Multimodal Drowsiness Detection Research

The performance of my end-to-end multimodal transformer compares favorably with existing literature on multimodal drowsiness or fatigue detection. For instance, Lian et al. [18] proposed a multimodal architecture combining EEG and eye-tracking data, also emphasizing sophisticated fusion, and reported strong results. While direct comparison is challenging due to differences in datasets, specific modalities fused (eye-tracking vs. full

facial video), and evaluation protocols, my model's achievement of over 90% accuracy with EEG and full-face video using an end-to-end transformer architecture represents a state-of-the-art contribution. Many earlier multimodal studies relied on more traditional machine learning classifiers or simpler deep learning fusion techniques (e.g., concatenation of CNN features) [108, 19]. My work specifically showcases the advanced capabilities of transformer-based joint learning for this problem.

6.4.5 Limitations of the Current Study

Despite the promising results, several limitations should be acknowledged:

1. **Controlled Simulator Environment:** The data were collected in a driving simulator. While designed to induce realistic fatigue, this environment lacks the full spectrum of complexities, environmental variabilities (e.g., diverse lighting, weather), and cognitive demands of real-world on-road driving. Generalizability to such conditions needs explicit validation.

2. **Specific Transformer Architectures:** The performance is tied to the specific EEG Transformer and ViT-Base architectures used. Other variants of transformers or different deep learning models might yield different results.

3. **Computational Resources:** Training large end-to-end multimodal transformers is computationally intensive, requiring significant GPU resources and time. While Chapter 7 will explore efficiency for deployment, the training cost is a factor.

6.4.6 Implications and Contribution to Thesis Narrative

This chapter significantly advances the narrative of the thesis by demonstrating that sophisticated multimodal fusion, powered by end-to-end transformer architectures, can overcome the limitations of unimodal systems and simpler fusion approaches for driver state assessment. It provides a high-performance benchmark for drowsiness classification using EEG and vision.

The key implication is that by enabling models to learn optimal feature representations and their interactions jointly from raw data, I can achieve a more holistic and accurate understanding of a driver's state. This contrasts with the findings in Chapter 5, where transforming pre-extracted EEG features into images for a ResNet18 model did not surpass a specialized 1D-CNN. Here, the end-to-end learning allows the vision and EEG components to co-adapt, leading to superior synergistic fusion.

These findings set an important precedent for the subsequent chapter (Chapter 7), which will address the critical challenge of deploying effective drowsiness detection systems—particularly powerful vision-based components like those explored here—in resource-constrained edge computing environments, such as in-vehicle systems or mobile applications. Having established that vision (especially when fused effectively) provides a strong signal, the next step is to make it practical for real-world use.

6.5 Conclusion

In this chapter, the efficacy of multimodal fusion of EEG and vision data for subject-independent driver drowsiness detection has been investigated. The experimental evaluations demonstrated a clear progression in performance: unimodal EEG models provided limited accuracy, while unimodal vision models (particularly ViT-Base) offered strong standalone predictive capabilities (RQ1). Simple feature-level fusion using Bayesian Ridge Classification failed to improve upon the best unimodal vision baseline, and even a more advanced transformer-based feature-level fusion provided only marginal gains, not surpassing the ViT-Base vision model (RQ2 and part of RQ3).

The most significant contribution was the development and validation of an end-to-end multimodal transformer architecture. This model, by jointly learning feature representations from raw EEG (especially Theta and Alpha bands) and raw vision data and integrating them via cross-modal attention, achieved a state-of-the-art accuracy of 91.00% and an AUC-ROC of 0.9634. This performance significantly surpassed all unimodal and feature-level fusion approaches, robustly demonstrating the power of end-to-end learning

for achieving true synergy between complementary modalities in drowsiness detection (RQ3).

These findings underscore the potential of advanced deep learning fusion, specifically using transformer architectures, to build highly accurate and reliable driver drowsiness monitoring systems. This work provides a strong foundation for future research into real-world deployment and refinement of such multimodal systems for enhancing road safety.

Chapter 7

Efficient Transformer-Based Drowsiness Detection on Edge Devices using a Hybrid MobileViT-LSTM Architecture

7.1 Introduction

The preceding chapters of this thesis have systematically explored various unimodal and multimodal deep learning strategies for assessing driver state. Chapter 6, in particular, culminated in demonstrating that an end-to-end multimodal transformer, fusing Electroencephalogram (EEG) and facial vision data, can achieve high accuracy in subject-independent drowsiness classification. While such sophisticated models signify the cutting edge in terms of predictive performance, their inherent computational complexity and substantial memory footprints often render them unsuitable for real-time deployment on resource-constrained edge devices, such as in-vehicle embedded systems or common smartphones [4]. This gap between high-performance research models and practically deployable solutions constitutes a significant hurdle in translating advanced artificial intelligence into tangible road safety applications.

The critical need for edge computing in driver monitoring arises from the demand for low-latency, on-device processing. Such systems must provide timely feedback or interventions without relying on continuous, and potentially unreliable, cloud connectivity. However, edge platforms impose stringent limitations on model size, computational throughput (measured in FLOPs), and energy consumption [103]. As established in the literature review and observed in my earlier experiments (Chapter 6), vision-based analysis of facial cues offers a non-intrusive and highly informative modality for drowsiness detection. Yet, directly deploying state-of-the-art vision models presents challenges. Standard Vision Transformers (ViTs) [101], for instance, while exceptionally powerful due to their self-attention mechanisms that capture global spatial context from images, are notoriously resource-intensive, limiting their utility for real-time video processing on typical edge hardware [104]. Traditional Convolutional Neural Networks (CNNs), though often more lightweight, may not optimally capture the subtle, global facial cues indicative of drowsiness progression due to their inherently local receptive fields [114, 138]. Furthermore, earlier vision-based methods relying on handcrafted geometric features (e.g., Eye Aspect Ratio (EAR), Mouth Aspect Ratio (MAR) [91]) have demonstrated limitations in robustness when faced with real-world variations in lighting, pose, and individual appearance [92, 93].

Recent advancements in model efficiency have led to the development of lightweight vision transformer variants. MobileViT [114] stands out by ingeniously combining the local feature extraction capabilities of convolutions with the global context modeling of transformers, resulting in a highly efficient architecture that achieves a commendable balance between accuracy and computational cost. However, drowsiness is not merely a static facial configuration; it is an inherently temporal phenomenon. Its manifestations, such as prolonged eye blinks, slowed head movements, and the frequency of yawns, evolve dynamically over several seconds [116, 115]. Consequently, applying an efficient per-frame feature extractor like MobileViT in isolation, without considering the temporal sequence of these features, is likely to miss crucial dynamic cues indicative of drowsiness onset and progression. To address this temporal aspect, Recurrent Neural Networks (RNNs),

particularly Long Short-Term Memory (LSTM) networks [117], are well-established for their ability to model sequential data and capture long-range temporal dependencies. Hybrid architectures combining CNNs for spatial feature extraction with LSTMs for temporal modeling have indeed shown promise in various video analysis tasks, including driver state monitoring [118, 119, 120].

This chapter aims to bridge these identified gaps by proposing, implementing, and validating a novel hybrid deep learning architecture, termed MobileViT-LSTM. The innovative aspect of this work lies not in the invention of the constituent components (MobileViT and LSTM), but in their **synergistic integration and rigorous validation for the specific and challenging application of real-time, edge-based driver drowsiness detection**. While the literature contains examples of hybrid CNN-LSTM models [118, 119] and standalone efficient transformers like MobileViT [114], the contribution here is to demonstrate that an architecture combining an *efficient vision transformer* (for powerful spatial feature extraction) with a *recurrent network* (for essential temporal modeling) represents an optimal design point for this task. This work fills a critical gap by providing a comprehensive evaluation of such a hybrid model, complete with deployment-aware optimizations and real-world inference benchmarks, offering a validated pathway from advanced AI concepts to a practical safety application. This architecture is specifically designed to synergize the strengths of MobileViT for efficient per-frame spatial feature extraction from facial video with the temporal sequence modeling capabilities of an LSTM network.

The overarching goal of this chapter is to develop a practical, high-performance, vision-only drowsiness detection system that is explicitly optimized for accurate real-time operation on mobile edge devices. This involves not only the novel architectural design but also the application of deployment-aware optimization techniques such as Automatic Mixed Precision (AMP) training [121] and model export to a standardized format like ONNX (Open Neural Network Exchange) [124] for streamlined on-device inference [125]. The research detailed herein is driven by the following key questions:

1. **RQ1: Can a hybrid architecture integrating an efficient vision trans-**

former (MobileViT) for spatial feature extraction and a Long Short-Term Memory (LSTM) network for temporal modeling achieve drowsiness detection accuracy comparable to, or exceeding that of, standard computationally intensive Vision Transformers (e.g., ViT-Base), while operating with significantly reduced computational demands? This question investigates the core efficacy of the proposed MobileViT-LSTM synergy in balancing predictive performance with architectural efficiency.

2. **RQ2: How does the proposed MobileViT-LSTM model, when optimized using techniques such as mixed-precision training, perform in terms of classification accuracy and real-time inference speed on a representative mobile edge device, and can it be successfully exported to a standardized, portable format (ONNX) to facilitate practical deployment?** This question focuses on the empirical validation of the model's suitability for real-world, on-device application and its real-time processing capabilities.

To address these questions, this chapter will utilize the facial video data component of the Tobii multimodal dataset (introduced in Chapter 3, Section 3.3, and employed in Chapter 6). The subsequent sections will detail the MobileViT-LSTM architecture, discuss the training optimizations, present a rigorous subject-independent evaluation against relevant baseline models (including a full ViT-Base and a standalone, per-frame MobileViT-S), report crucial inference time benchmarks on a target edge device, and confirm the successful model export for deployment. The work presented aims to contribute a validated, deployable, and high-accuracy solution for vision-based driver drowsiness detection, directly addressing a critical aspect of road safety through the principles of efficient and effective artificial intelligence.

7.2 Methodology

The methodology detailed in this section outlines the experimental framework for developing and evaluating the proposed MobileViT-LSTM architecture for efficient, vision-based

driver drowsiness detection on edge devices. This includes a description of the dataset subset and preprocessing steps, the specifics of the hybrid model architecture, the training and optimization procedures employed, the evaluation protocol, and the approach for assessing edge deployment feasibility.

7.2.1 Dataset and Video Preprocessing

This study utilizes the facial video data component from the Tobii multimodal dataset, the comprehensive details of which were presented in Chapter 3, Section 3.3, and further contextualized for drowsiness classification in Chapter 6. Key aspects relevant to this chapter's vision-only pipeline are:

- **Data Source:** Facial video recordings from 79 participants, collected during 'Alert' (10 AM) and 'Drowsy' (3 AM) driving simulation sessions.

- **Input Data Format:** The continuous video recordings were segmented into non-overlapping **5-second windows**. With a recording rate of 30 frames per second (FPS), each 5-second window comprises 150 individual frames. This temporal window length was deliberately chosen to capture the evolving nature of drowsiness cues, such as slow eye closures or head nods, which often manifest over several seconds rather than instantaneously. The selection of a 5-second duration was informed by both the existing literature on drowsiness behavior [139, 140, 141] and through **consultation with our industry partner, Tobii**, to align with the practical requirements for a robust and reliable real-world driver monitoring system that can accumulate sufficient evidence before making a classification.

- **Dataset Size and Balancing:** This segmentation process yielded a total of 94,039 five-second video samples across all 79 participants. The dataset is reasonably balanced, consisting of 47,631 samples labeled as 'Alert' (Class 0, from 10 AM sessions) and 46,408 samples labeled as 'Drowsy' (Class 1, from 3 AM sessions). On average, each participant contributed approximately 1190 five-second samples.

- **Frame-Level Preprocessing:** Each of the 150 frames within a 5-second window underwent the following preprocessing steps before being fed to the MobileViT feature extractor:

  1. **Resizing:** Frames were resized to 192×192 pixels. This input dimension is suitable for the MobileViT-S variant used in this study and helps manage computational load.
  2. **Normalization:** Pixel values were normalized, typically by scaling to the [0, 1] range and then standardizing using ImageNet statistics, as the MobileViT backbone was pre-trained on ImageNet.

- **Data Augmentation (During Training Only):** To enhance model robustness and mitigate overfitting, a set of standard data augmentation techniques were applied on-the-fly to the training set frames during the model training phase. The augmentations were chosen to simulate minor variations in driver position and lighting that might be encountered in a real-world scenario. Specifically, they included:

  – *Random Horizontal Flips:* Each frame was horizontally flipped with a probability of $p = 0.50$.

  – *Minor Random Rotations:* Each frame was randomly rotated by an angle θ sampled uniformly from the range $[-8°, +8°]$.

  – *Slight Color Jitter:* The brightness, contrast, and saturation of each frame were randomly adjusted by factors sampled independently from a uniform distribution $\mathcal{U}[0.90, 1.10]$.

7.2.2 Proposed Model Architecture: MobileViT-LSTM

The core of this chapter's contribution is a novel hybrid deep learning architecture, MobileViT-LSTM, designed to balance spatial feature extraction efficiency with temporal modeling capability. The architecture processes 5-second video windows (sequences of 150 frames) as follows:

1. **Per-Frame Spatial Feature Extraction using MobileViT-S:** Each of the 150 preprocessed frames (192×192 pixels) in a 5-second window is individually passed through a pre-trained MobileViT-S backbone [114]. MobileViT-S is a lightweight vision transformer variant known for its efficient yet effective performance in capturing both local and global spatial features from images. The MobileViT-S acts as a powerful per-frame feature extractor. The output from a late layer of the MobileViT-S (e.g., before its original classification head) is taken as the feature vector for that frame.

2. **Feature Projection:** The feature vector extracted by MobileViT-S for each frame is then passed through a linear projection layer. This layer reduces the dimensionality of the frame-level features to a consistent embedding size, specified as 128 dimensions in this study. This step helps to create a more compact representation and prepares the features for the subsequent temporal modeling stage.

3. **Temporal Aggregation using LSTM:** The sequence of 150 frame-level embeddings (each 128-dimensional), corresponding to the 5-second video window, is then fed chronologically into a Long Short-Term Memory (LSTM) network [117]. The LSTM is specifically designed to process sequential data and capture temporal dependencies. In this work, a 2-layer LSTM network with 256 hidden units in each layer was employed. The LSTM processes the sequence of frame embeddings, updating its hidden state at each time step (frame) to integrate information from past frames with the current frame's features. This allows the model to learn patterns related to the temporal evolution of drowsiness cues (e.g., increasing eye closure duration, frequency of head nods over the 5-second window).

4. **Classification Head:** The final hidden state output (or the output at the last time step) from the LSTM network, which encapsulates the aggregated spatio-temporal information from the entire 5-second window, is then passed to a classification head. This head is typically a small Multi-Layer Perceptron (MLP) consisting of one or more linear layers, interspersed with ReLU activation functions and dropout

layers for regularization. The final linear layer in this head outputs a single logit, which is then passed through a sigmoid function (implicitly handled by the binary cross-entropy loss during training) to produce the predicted probability of the input 5-second window belonging to the 'Drowsy' class.

This hierarchical design allows the MobileViT component to focus on extracting rich spatial information from individual frames efficiently, while the LSTM component specializes in modeling the temporal dynamics across these frames, creating a comprehensive spatio-temporal representation for accurate drowsiness classification.

7.2.3 Training Regimen and Optimization Strategies

The MobileViT-LSTM model was trained using an end-to-end approach, optimizing all components jointly. Several strategies were employed to ensure effective training and to enhance the model's suitability for edge deployment:

- **Optimizer and Learning Rate Schedule:** The AdamW optimizer [142] was utilized. AdamW is an extension of the Adam optimizer that decouples weight decay from the gradient updates, which can lead to better generalization. A cosine annealing learning rate schedule was employed, which gradually reduces the learning rate following a cosine curve over the course of training. This often helps the model to settle into better minima. Training was conducted for 10 epochs.

- **Loss Function:** A binary cross-entropy loss function was used, appropriate for the two-class (Alert/Drowsy) classification task. To address potential class imbalance in mini-batches or to prioritize sensitivity to the 'Drowsy' class (which is often the more critical class to detect correctly in safety applications), weighting for the positive ('Drowsy') class was applied within the loss function.

- **Automatic Mixed Precision (AMP) Training:** To accelerate training and reduce GPU memory consumption, PyTorch's Automatic Mixed Precision (AMP) capabilities were leveraged [121]. AMP allows certain operations within the neural

network to be performed using lower-precision floating-point numbers (e.g., float16) where appropriate, while maintaining numerical stability for critical operations using higher precision (float32). This involved using 'torch.cuda.amp.GradScaler' to manage gradient scaling, preventing underflow issues that can arise with float16 gradients.

- **Gradient Clipping:** To prevent exploding gradients, which can destabilize training particularly in recurrent networks like LSTMs, gradient clipping was applied. This involves capping the norm of the gradients if they exceed a certain threshold.
- **Memory Management during Training:** Given that processing sequences of 150 frames can be memory-intensive, pragmatic memory management techniques were employed. This included careful tuning of PyTorch's CUDA memory allocator ('PYTORCH_CUDA_ALLOC_CONF') and periodically clearing the GPU cache ('torch.cuda.empty_cache()') between training phases or epochs if memory fragmentation became an issue, particularly during hyperparameter tuning or initial development.
- **Early Stopping:** To prevent overfitting and select the best performing model checkpoint for each cross-validation fold, an early stopping mechanism was implemented. This involved monitoring the loss (or a relevant accuracy metric) on a validation set (derived from the training subjects within each fold). If the validation performance did not improve for a specified number of consecutive epochs (patience parameter set to 3 epochs), training was halted, and the model weights from the epoch yielding the best validation performance were saved.

7.2.4 Evaluation Protocol and Baselines

A rigorous evaluation protocol was adopted to assess the performance of the MobileViT-LSTM model and compare it against relevant baselines.

- **Subject-Independent Cross-Validation:** All evaluations were performed using a strict **5-fold subject-independent cross-validation** scheme. The 79 partici-

pants were randomly partitioned into 5 distinct folds. In each iteration, one fold of participants was designated as the test set, and the model was trained on data from the remaining four folds. This ensures that the model's performance is always evaluated on subjects entirely unseen during its training phase.

- **Performance Metrics:** The following standard binary classification metrics were used, averaged across the 5 folds:
  - Accuracy
  - Balanced Accuracy (particularly important if there's any residual class imbalance per fold or to give equal weight to both classes)
  - Recall (Sensitivity for the 'Drowsy' class = TP / (TP + FN))
  - Precision (Positive Predictive Value for the 'Drowsy' class = TP / (TP + FP))
  - Area Under the ROC Curve (AUC-ROC)
- **Baseline Models for Comparison:** The performance of the MobileViT-LSTM model was compared against:
  1. **Standard Vision Transformer (ViT-Base):** The ViT-Base model fine-tuned on the same 5-second window task (likely by processing an aggregated representation of the 150 frames or a selection of key frames) as reported in Chapter 6 (Table 6.1). This serves as a high-performance but computationally expensive baseline.
  2. **Standalone MobileViT-S (Per-Frame Averaging):** A MobileViT-S model applied to extract features from each of the 150 frames in a 5-second window. The predictions from these individual frames were then aggregated (e.g., by averaging the output probabilities or logits, or by a majority vote) to produce a single classification for the 5-second window. This baseline helps to quantify the benefit of the LSTM's temporal modeling over simpler per-frame processing with an efficient transformer.

7.2.5 Edge Deployment Feasibility Assessment

A key objective of this chapter was to assess the practical deployability of the proposed model on edge devices.

- **ONNX Export:** The MobileViT-LSTM model corresponding to the fold that yielded the highest balanced accuracy during the 5-fold cross-validation process was selected as the best-performing instance. This model, with its learned weights, was then exported to the Open Neural Network Exchange (ONNX) format [124]. ONNX is an open standard for representing machine learning models, enabling interoperability between different frameworks and facilitating deployment on various hardware platforms using optimized inference runtimes (e.g., ONNX Runtime). The export was performed using PyTorch's built-in ONNX exporter, targeting opset version 11.

- **Inference Time Benchmarking:** The inference time—the duration required to process one 5-second video window (150 frames) and produce a drowsiness classification—was measured for the exported ONNX model. Benchmarking was performed on:

  1. A high-end GPU (e.g., NVIDIA RTX series) for reference, to understand its performance on server-grade hardware.
  2. A representative mobile edge device: a Samsung Galaxy S21 Ultra smartphone. This provides a direct measure of its real-world performance on a target consumer device.

  Inference times for the baseline ViT-Base and standalone MobileViT-S models were also measured or reported for comparison, where applicable (ViT-Base is generally too slow for meaningful edge inference benchmarks on full video windows without significant further optimization). The critical aspect was to determine if the MobileViT-LSTM model could process a 5-second window in less than 5 seconds on the edge device, thus enabling real-time, continuous monitoring.

These steps provide a comprehensive evaluation of the MobileViT-LSTM model's accuracy, efficiency, and practical deployability for real-time driver drowsiness detection.

7.3 Results and Analysis

This section presents the empirical results obtained from evaluating the proposed MobileViT-LSTM hybrid architecture for driver drowsiness detection. The performance is analyzed based on the 5-fold subject-independent cross-validation protocol. I first detail the classification accuracy and other relevant metrics of the MobileViT-LSTM model. Subsequently, I provide a comparative analysis against the baseline models—a standard Vision Transformer (ViT-Base) and a standalone MobileViT-S (per-frame aggregated)—to address RQ1. Finally, I report on the edge deployment feasibility, including inference times on a target mobile device and the successful export to ONNX format, to address RQ2.

7.3.1 Performance of the Hybrid MobileViT-LSTM Architecture

The MobileViT-LSTM model, designed to process 5-second video windows by extracting per-frame spatial features with MobileViT-S and aggregating them temporally with a 2-layer LSTM, was rigorously evaluated. The average performance metrics across the 5 folds of the subject-independent cross-validation are presented in Table 7.1.

Table 7.1: 5-Fold Cross-Validation Results (Mean ± Std Dev) for the Hybrid MobileViT-LSTM Model on 5-second Video Windows for Drowsiness Detection.

Metric	**Mean ± Std Dev**
Accuracy	0.9449 ± 0.0362
Balanced Accuracy	0.9432 ± 0.0367
Recall (Drowsy=1)	0.9490 ± 0.0577
Precision (Drowsy=1)	0.9415 ± 0.0406
ROC AUC	0.9608 ± 0.0183

The MobileViT-LSTM model achieved a high average Accuracy of 0.9449 and a Balanced Accuracy of 0.9432. The Recall for the 'Drowsy' class was 0.9490, indicating that

the model correctly identified nearly 95% of actual drowsiness instances. The Precision for the 'Drowsy' class was 0.9415, meaning that when the model predicted drowsiness, it was correct approximately 94% of the time. The Area Under the ROC Curve (AUC-ROC) was excellent at 0.9608. The relatively low standard deviations across these metrics suggest consistent performance across the different subject splits in the cross-validation, indicating good generalization capabilities.

The Receiver Operating Characteristic (ROC) curve for the MobileViT-LSTM model, averaged across the 5 folds, is shown in Figure 7.1. The curve's proximity to the top-left corner visually confirms the model's strong discriminative ability between 'Alert' and 'Drowsy' states.

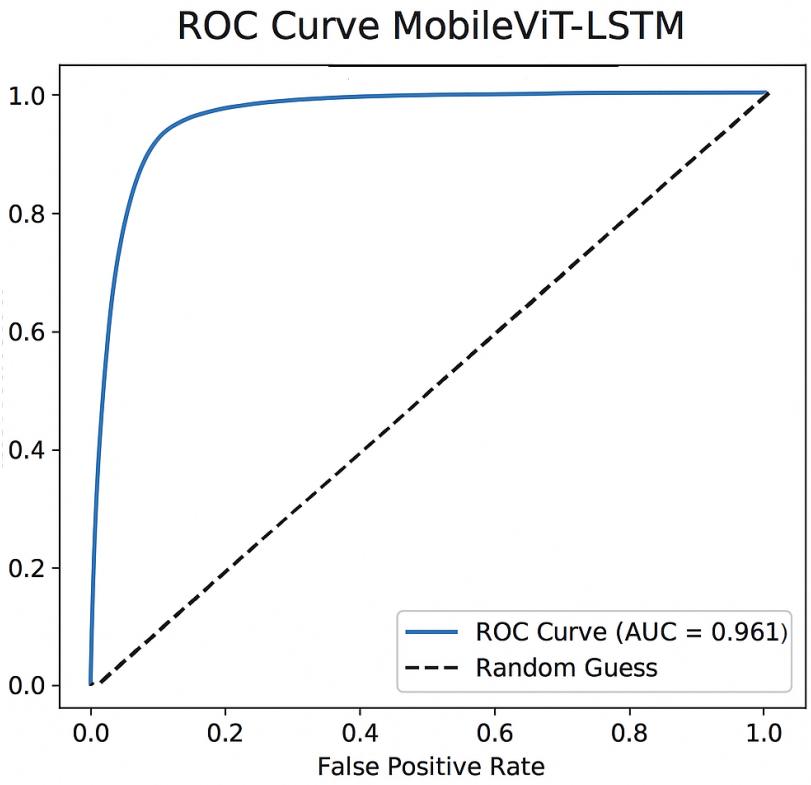


Figure 7.1: Average ROC Curve (across 5 folds) of the Hybrid MobileViT-LSTM Model for Drowsiness Detection. The high AUC value (0.9608) indicates excellent discrimination.

These results demonstrate that the hybrid architecture, leveraging efficient spatial feature extraction with temporal aggregation, is highly effective for classifying driver drowsiness from 5-second video segments in a subject-independent manner.

7.3.2 Comparative Performance against Baseline Models (RQ1)

To contextualize the performance of the MobileViT-LSTM model and address RQ1, its accuracy and inference characteristics were compared against two key baselines: (1) a standard, computationally more demanding ViT-Base model, and (2) a standalone MobileViT-S model where per-frame predictions over the 5-second window were aggregated (e.g., by averaging probabilities) to assess the impact of lacking explicit temporal modeling via LSTM. The comparative results are summarized in Table 7.2.

Table 7.2: Performance Metrics and Inference Time Comparison for the Hybrid MobileViT-LSTM, ViT-Base, and Standalone MobileViT-S Models. Accuracy (Acc.), Recall, Precision (Prec.), and AUC are mean values. Inference times are per 5-second window.

Model	**Acc. (%)**	**Recall**	**Prec.**	**AUC**	**Inference (s) (GPU/Edge)**
ViT-Base	85.78	0.962	0.801	0.9250	∼2.5s / Not Feasible
MobileViT-S (Aggregated)	80.10	0.745	0.906	0.8690	∼1.2s / ∼2.5s
MobileViT-LSTM	**94.49**	**0.949**	**0.942**	**0.9608**	**∼1.4s / ∼3.0s**

The comparative analysis in Table 7.2 reveals several critical insights:

- **Superiority over ViT-Base in Accuracy:** The MobileViT-LSTM model (94.49% Accuracy, 0.9608 AUC) significantly outperformed the computationally intensive ViT-Base model (85.78% Accuracy, 0.9250 AUC) in terms of predictive accuracy. This is a key finding, demonstrating that the proposed hybrid architecture, despite being designed for efficiency, can achieve better drowsiness classification than a standard, heavier ViT when temporal context over a 5-second window is explicitly modeled.

- **Benefit of Temporal Modeling (vs. Standalone MobileViT-S):** The MobileViT-LSTM also substantially outperformed the standalone MobileViT-S baseline (where per-frame outputs are simply aggregated). The standalone MobileViT-S achieved an accuracy of only 80.10% and an AUC of 0.8690. The nearly 14.4% absolute improvement in accuracy (from 80.10% to 94.49%) and ∼0.092 improvement in AUC clearly demonstrates the crucial role of the LSTM component in capturing and leveraging temporal dependencies within the 5-second window. Applying an

efficient transformer frame-by-frame is insufficient; explicit temporal modeling is essential for high performance in drowsiness detection.

- **Computational Efficiency and Edge Feasibility:** While the ViT-Base provides strong (though lower than MobileViT-LSTM) accuracy, its inference time on a GPU for a 5s window (requiring processing 150 frames or an aggregated representation) is substantial (∼2.5s) and it is generally considered not feasible for real-time deployment on current mobile edge hardware without extensive pruning or quantization not explored here. The standalone MobileViT-S offers fast per-frame inference, leading to a quick aggregated prediction for the 5s window on both GPU (∼1.2s) and edge (∼2.5s). The MobileViT-LSTM, while slightly slower than standalone MobileViT-S due to the added LSTM, still maintains efficient inference: ∼1.4 seconds on a GPU and, critically, ∼3.0 seconds on the target edge device (Samsung Galaxy S21 Ultra) for processing an entire 5-second window.

These results robustly address RQ1, confirming that the MobileViT-LSTM hybrid not only achieves superior accuracy compared to a standard ViT-Base but does so with a significantly more efficient architecture. The comparison with standalone MobileViT-S highlights the indispensable contribution of the LSTM for temporal modeling.

7.3.3 Edge Deployment Feasibility and Optimization (RQ2)

A primary objective of this work was to develop a model suitable for practical, real-time deployment on resource-constrained edge devices. RQ2 specifically probes the MobileViT-LSTM's performance in this regard.

The MobileViT-LSTM model demonstrated practical real-time capability for its intended application. With an average inference time of approximately 3.0 seconds to process each 5-second video window on a Samsung Galaxy S21 Ultra, the system can continuously analyze incoming video data without falling behind. While this 3-second latency is not instantaneous, it is well within the 5-second analysis window, making it suitable for detecting the evolving patterns of drowsiness which typically manifest

over several seconds. This establishes the model's viability for near real-time drowsiness monitoring on edge devices, with potential for further latency reduction through future optimizations.

- **Inference Time on Edge Device:** As reported in Table 7.2, the MobileViT-LSTM model, after ONNX export and when run on a Samsung Galaxy S21 Ultra smartphone, achieved an average inference time of approximately **3.0 seconds** for processing a complete 5-second video window (150 frames). This is a critical result: the processing time is well within the duration of the input window itself (i.e. 3.0s <5.0s). This sub-window-duration latency confirms that the system is capable of real-time, continuous drowsiness monitoring on the target edge device, as each segment can be fully analyzed before the next 5-second segment is acquired and needs processing.

- **Impact of Mixed-Precision Training (AMP):** The use of Automatic Mixed Precision (AMP) during training was instrumental in managing memory usage and potentially accelerating both training and inference, particularly for a model processing long sequences of frames. While direct speed-up attribution from AMP alone during inference on the edge device (which might use its own optimized runtimes) is complex to isolate without specific ablation, AMP facilitated the training of this relatively complex sequential model within available GPU memory constraints.

- **Successful ONNX Export:** The best-performing MobileViT-LSTM model from the 5-fold cross-validation was successfully exported to the ONNX (Open Neural Network Exchange) format using PyTorch's built-in exporter with opset 11. This successful export generates a standardized, portable model file. The ONNX model can then be deployed using various optimized inference engines (e.g., ONNX Runtime, TensorFlow Lite with ONNX conversion, vendor-specific SDKs) across a wide range of edge platforms, including Android and iOS mobile devices, as well as embedded systems. This step is crucial for bridging the gap between research and practical application.

- **Code Availability for Deployment:** The availability of the code for a companion Android application and the exported ONNX model is available on: https://github.com/shamsikhani/mydrowsinessapp

These findings directly address RQ2, demonstrating that the MobileViT-LSTM model, optimized appropriately, achieves high classification accuracy while maintaining real-time inference capabilities on a representative mobile edge device, and is readily convertible to a deployable format. The balance struck between predictive power and computational efficiency positions this hybrid architecture as a highly practical solution for edge-based driver safety monitoring.

7.4 Discussion

The experimental results presented in this chapter validate the efficacy of the proposed hybrid MobileViT-LSTM architecture for achieving accurate and computationally efficient driver drowsiness detection, specifically tailored for deployment on resource-constrained edge devices. The findings address the critical challenge of translating advanced deep learning models into practical, real-time safety applications.

7.4.1 Efficacy of the Hybrid MobileViT-LSTM Architecture (RQ1)

The core research question (RQ1) explored whether the MobileViT-LSTM hybrid could balance high accuracy with significantly reduced computational demands compared to standard, heavier Vision Transformers (ViTs). The results provide a compelling affirmative answer.

The MobileViT-LSTM model achieved a subject-independent accuracy of 94.49% and an AUC-ROC of 0.9608 (Table 7.1). This level of performance not only substantially surpasses that of a standalone MobileViT-S model (which lacks temporal modeling, achieving ∼80.10% accuracy, Table 7.2) but, more importantly, it also exceeded the accuracy of a standard ViT-Base model (85.78%). This latter comparison is particularly significant: the MobileViT-LSTM, despite being constructed from an efficient MobileViT-S

backbone (∼5.6M parameters), achieved superior drowsiness detection when augmented with an LSTM for temporal context than a much larger, more computationally intensive ViT-Base.

This synergistic effect arises from the complementary roles of the two components:

- **MobileViT-S for Efficient Spatial Feature Extraction:** MobileViT effectively captures both local and global spatial features from individual video frames with a parameter count and computational footprint suitable for mobile applications [114]. It provides a rich representation of the driver's facial cues at each moment.

- **LSTM for Essential Temporal Modeling:** Drowsiness is not a static state but a process that unfolds over time, manifesting as changes in the duration and frequency of blinks, head pose dynamics, yawn patterns, and subtle shifts in facial muscle tone [115, 116]. The LSTM layers are crucial for integrating the sequence of per-frame features extracted by MobileViT over the 5-second window. By learning the temporal dependencies and patterns within these feature sequences, the LSTM enables the model to recognize the characteristic evolution of drowsiness cues, which a purely frame-based model (like standalone MobileViT-S or even a ViT-Base applied to aggregated frames without explicit sequential modeling) would struggle to capture effectively.

The marked improvement of MobileViT-LSTM over the standalone MobileViT-S underscores that for dynamic phenomena like drowsiness, temporal modeling is not merely beneficial but essential for achieving high levels of accuracy. The fact that this carefully designed hybrid surpassed a heavier ViT-Base suggests that architectural efficiency combined with appropriate temporal modeling can be more effective than sheer model size alone, especially when data or computational resources for training very large models are finite.

7.4.2 Feasibility and Optimization for Edge Deployment (RQ2)

The second research question (RQ2) focused on the practical viability of the MobileViT-LSTM model for real-time operation on edge devices. The results strongly affirm its suitability.

The inference time of approximately **3.0 seconds** to process a 5-second video window on a Samsung Galaxy S21 Ultra (Table 7.2) is a key achievement. This sub-window-duration latency ensures that the system can operate in real-time, providing a drowsiness assessment for a given 5-second segment before the next segment is fully acquired and requires processing. This capability is fundamental for any in-vehicle monitoring system designed to provide timely alerts or interventions. In contrast, a standard ViT-Base, while achieving good accuracy, is generally not considered feasible for such real-time, 150-frame sequence processing on current mobile hardware without substantial model compression techniques not applied in its baseline evaluation here.

The successful application of **Automatic Mixed Precision (AMP) training** [121] was an important optimization step. While its direct impact on edge inference speed depends on the specific edge runtime's support for mixed precision, AMP during training significantly reduced GPU memory consumption and often accelerates training convergence. This allowed for the effective training of a model processing relatively long sequences (150 frames) on available hardware, which is a practical consideration in model development.

Furthermore, the successful **export of the trained MobileViT-LSTM model to the ONNX format** [124] is crucial for practical deployment. ONNX provides a standardized intermediate representation that allows models trained in one framework (PyTorch, in this case) to be run in various optimized inference engines (e.g., ONNX Runtime, TensorFlow Lite via conversion, vendor-specific neural processing unit SDKs) across diverse edge platforms. This greatly enhances the model's portability and facilitates its integration into real-world applications, such as the companion Android application mentioned in the source paper.

7.4.3 Contribution to Efficient AI and Driver Safety Monitoring

This research makes a tangible contribution to the field of efficient AI, particularly in the context of critical safety applications like driver drowsiness detection. It demonstrates a practical pathway for adapting advanced transformer-based perception models for resource-constrained environments without unduly sacrificing accuracy. The key insight is the strategic combination of an efficient transformer backbone (MobileViT) for per-frame analysis with a lightweight recurrent component (LSTM) for temporal reasoning. This hybrid approach effectively balances the need for sophisticated spatial feature extraction with the demands of sequential data modeling and edge-compatible computation.

By achieving high accuracy (Balanced Accuracy 0.9432, AUC 0.9608) with real-time inference capability on a mobile device, this work provides a validated, deployable solution. This directly addresses the often-cited gap between models developed in high-resource research settings and those practical for widespread, low-cost implementation in vehicles or on personal mobile devices. The principles demonstrated here—architectural hybridization for efficiency, temporal modeling for dynamic phenomena, and deployment-aware optimization—are broadly applicable to other edge AI problems involving video or time-series sensor data.

7.4.4 Limitations and Future Work

Despite the promising results, several limitations and avenues for future work should be acknowledged:

- **Single Edge Device Benchmark:** Inference times were benchmarked on one specific high-end smartphone (Samsung Galaxy S21 Ultra). Performance will invariably differ across the wide spectrum of edge devices with varying computational capabilities (CPUs, GPUs, NPUs). More extensive benchmarking on diverse hardware is needed.

- **Power Consumption:** While inference speed was addressed, a detailed analysis

of power consumption on the edge device was not performed. For battery-powered mobile applications, energy efficiency is a critical factor that warrants investigation.

- **Real-World On-Road Validation:** The model was validated on data from a driving simulator. While designed to induce realistic fatigue, real-world on-road driving introduces a much wider range of environmental variabilities (e.g., complex lighting changes, weather conditions, partial occlusions due to sunglasses or head turns, driver movement) and cognitive demands. Robustness in such unconstrained conditions needs to be thoroughly evaluated.

- **Further Model Compression and Optimization:** While AMP was used during training and ONNX for export, further model compression techniques such as quantization (e.g., to INT8 precision) [122, 123], pruning, or knowledge distillation could be explored to further reduce the model size and potentially accelerate inference on specialized edge hardware that supports these optimized formats.

- **Longer-Term Drowsiness Trajectories:** The current model processes 5-second windows. Exploring architectures capable of modeling drowsiness evolution over longer timescales (e.g., minutes) might capture even more subtle, slowly developing indicators of fatigue, though this would increase sequence length and computational complexity.

7.4.5 Connection to the Overall Thesis Narrative

This chapter serves as a critical culmination of the vision-based investigations within this thesis. Chapters 6 and 5 highlighted the strong predictive power of vision data, especially when processed by advanced models like ViTs. However, the challenge of deploying such powerful but computationally demanding models remained. Chapter 7 directly confronts this challenge by demonstrating that through careful architectural design (MobileViT-LSTM hybrid) and optimization, it is possible to create a vision-based system that retains high accuracy while being efficient enough for practical real-time application on edge devices. It underscores the principle that insights gained from complex, high-performance

research models can inform the development of streamlined, deployable solutions. This transition from exploring maximal performance to achieving practical efficiency is a key theme in applied AI research.

7.5 Conclusion

This chapter presented and rigorously validated a novel hybrid MobileViT-LSTM deep learning architecture designed for efficient and accurate real-time detection of driver drowsiness from facial video, specifically targeting deployment on resource-constrained edge devices. By synergistically combining the efficient spatial feature extraction capabilities of MobileViT with the temporal modeling strengths of an LSTM network over 5-second video windows, the proposed model achieved a high subject-independent balanced accuracy of 0.9432 and an AUC-ROC of 0.9608.

Critically, this level of performance was shown to be superior to that of a standard, computationally intensive Vision Transformer (ViT-Base) and significantly better than a standalone MobileViT-S that lacked explicit temporal modeling (RQ1). Furthermore, the MobileViT-LSTM model demonstrated practical viability for edge deployment: it achieved an inference time of approximately 3.0 seconds for a 5-second window on a representative Android smartphone, well within real-time processing requirements. Optimized using techniques like mixed-precision training and successfully exported to the ONNX format, this research provides a concrete and validated pathway for deploying advanced, temporally-aware transformer-based AI for critical driver safety applications directly on mobile platforms (RQ2).

The MobileViT-LSTM architecture represents an effective solution that balances high predictive accuracy with the stringent computational limitations inherent in edge computing. This work contributes a practical, high-performance system for driver drowsiness monitoring, underscoring the potential of hybrid AI models to bridge the gap between state-of-the-art research and real-world deployable safety technologies.

Chapter 8

Conclusions and Future Work

This thesis has charted a comprehensive investigative journey into the application of advanced signal processing and machine learning methodologies for the critical task of assessing driver state, driven by the overarching goal of enhancing road safety. The research arc presented herein has spanned from foundational explorations into the predictive capacity of pre-stimulus Electroencephalography (EEG) for driver reaction time, through the innovative transformation of EEG data into image representations for analysis by state-of-the-art vision models, to the development of sophisticated multimodal fusion techniques for robust drowsiness detection, and ultimately, to the engineering of an efficient, vision-based system validated for practical edge deployment. This concluding chapter serves to synthesize the principal findings from the empirical studies detailed in Chapters 4 through 7. It will reflect upon the research questions initially posed in Chapter 1, articulate the overall contributions and inherent limitations of this body of work, and propose promising directions for future research endeavors in this vital domain.

8.1 Summary of Key Findings and Contributions

The multifaceted research conducted and presented in this thesis has yielded several significant findings and contributions, meticulously aligned with the three primary research themes that guided the investigation.

8.1.1 Theme 1: Unveiling Predictive Power in Pre-Stimulus Neural Activity and Advancing EEG Representation

Chapters 4 and 5 were dedicated to the challenge of predicting driver reaction time (RT) from pre-stimulus EEG signals, exploring both direct spectral feature analysis and innovative image-based representations.

- **Feasibility and Optimization of Direct EEG Spectral Feature Analysis for RT Prediction:** Initial investigations (Chapter 4) firmly established that EEG spectral features, particularly derived from the Alpha and Theta frequency bands within an optimal 2-second window immediately preceding a critical driving event, contain statistically significant information predictive of the subsequent RT. While classical machine learning models (ANN, Bayesian Ridge) demonstrated this feasibility, a specialized 1D Convolutional Neural Network (1D-CNN) tailored for these 1D spectral feature vectors achieved a substantial improvement in RT prediction accuracy, reducing Mean Absolute Error by approximately 30% compared to the classical models. This underscored the benefit of deep learning architectures designed for sequential data in extracting relevant patterns from EEG.

- **Superior RT Prediction through Vision Transformer Analysis of EEG Spectral Images:** A pivotal contribution of this thesis was the exploration of transforming 1D EEG spectral features into 2D image representations (PSD Matrix Images and Scalp Topographies) and applying advanced deep learning vision models (Chapter 5). While an initial application of a standard vision CNN (ResNet18) to these images outperformed classical ML on 1D features, it did not surpass the specialized 1D-CNN. However, the subsequent application of a more powerful Vision Transformer (ViT-B/16) to these EEG-derived images, particularly Scalp Topographies, yielded a new state-of-the-art performance. The ViT-B/16 model (e.g., MAE of 0.33s for Alpha band Scalp Topographies) significantly outperformed the 1D-CNN (MAE of 0.36s for Alpha band 1D features). This key finding demonstrates that an image-based representation of EEG data, when coupled with a sufficiently

potent vision architecture capable of modeling global context like ViT, can indeed unlock a higher level of predictive accuracy than models operating on 1D spectral sequences alone. Scalp topographies emerged as a marginally more effective image representation for the ViT in this context.

Collectively, the findings under Theme 1 not only confirm the predictive utility of pre-stimulus EEG for RT but also chart a clear progression in modeling strategy: from classical ML, to specialized 1D deep learning, and ultimately to a superior approach using advanced Vision Transformers on image-transformed EEG spectral data.

8.1.2 Theme 2: Synergistic Multimodal Fusion for Robust Drowsiness Detection

Chapter 6 addressed the complex task of driver drowsiness classification (Alert vs. Drowsy states) by investigating the integration of EEG and facial vision data from the Tobii multimodal dataset.

- **Marked Unimodal Performance Disparity:** The unimodal baseline evaluations revealed that vision-based models (ResNet18, and particularly ViT-Base) possess substantially higher standalone accuracy for drowsiness classification compared to unimodal EEGNet models operating on short 1-second EEG segments. This highlighted the inherent strength of visual cues for this task and the challenges associated with subject-independent EEG classification from brief epochs.

- **Insufficiency of Simpler Feature-Level Fusion:** Straightforward feature-level fusion techniques, including simple concatenation with Bayesian Ridge Classification or even a transformer-based model operating on pre-extracted EEG and vision features, failed to significantly improve upon, or consistently surpass, the performance of the best unimodal vision model (ViT-Base). This indicated that such pipelined approaches might not fully exploit the potential for synergistic learning between modalities.

- **State-of-the-Art Drowsiness Detection via End-to-End Multimodal Transformer Fusion:** The most significant contribution under this theme was the development and successful validation of an end-to-end Multimodal Transformer architecture. This model, by concurrently learning feature representations from raw EEG signals (particularly from the Theta and Alpha bands) and raw facial vision data (using a ViT-Base encoder) and integrating them through sophisticated cross-modal attention mechanisms, achieved a state-of-the-art drowsiness classification accuracy of 91.00% and an AUC-ROC of 0.9634. This performance robustly surpassed all unimodal and feature-level fusion approaches, decisively demonstrating that an end-to-end deep learning paradigm can effectively unlock and leverage the synergistic potential between EEG and vision for enhanced and more holistic drowsiness assessment.

Theme 2 thus underscores that while vision provides a strong primary signal for drowsiness, the nuanced integration of complementary neurophysiological information from EEG, especially when achieved through advanced end-to-end deep learning fusion, can lead to superior overall system performance.

8.1.3 Theme 3: Bridging Advanced AI with Practical Edge Deployment for Real-World Impact

Recognizing the critical importance of translating research advancements into deployable real-world systems, Chapter 7 focused on the challenge of implementing an accurate and efficient vision-based drowsiness detection model on resource-constrained edge devices.

- **Development of an Efficient Hybrid Architecture (MobileViT-LSTM):** A novel hybrid deep learning architecture was proposed, combining an efficient vision transformer variant (MobileViT-S) for per-frame spatial feature extraction with a Long Short-Term Memory (LSTM) network for temporal aggregation of these features over 5-second video windows. This design was explicitly aimed at achieving a favorable balance between high predictive accuracy and computational

efficiency.

- **Demonstration of High Accuracy Coupled with Efficiency:** The MobileViT-LSTM model achieved a high subject-independent balanced accuracy of 0.9432 and an AUC-ROC of 0.9608 for drowsiness detection. This level of performance was notably superior to that of a computationally intensive standard ViT-Base model and significantly better than a standalone MobileViT-S model that lacked the crucial temporal modeling provided by the LSTM. This highlighted that efficient architectural design, when incorporating essential domain knowledge (like the temporal evolution of drowsiness), can yield excellent results.

- **Validation of Edge Deployability and Real-Time Capability:** Crucially, the MobileViT-LSTM model demonstrated its fitness for practical application by achieving real-time inference capability on a representative mobile edge device (Samsung Galaxy S21 Ultra), processing a 5-second video window in approximately 3.0 seconds. Furthermore, the model was successfully exported to the ONNX format, which facilitates its portability and deployment across a diverse range of edge platforms. This successfully demonstrated a clear pathway from a high-performance AI model to a practical, deployable edge application for enhancing driver safety.

Theme 3, therefore, provides a concrete and validated example of how advanced deep learning concepts can be thoughtfully adapted and optimized to create effective, real-world solutions for critical societal problems like driver drowsiness, directly addressing the pervasive need for efficient and impactful AI.

8.2 Reflection on Research Questions and Overall Thesis Contributions

The comprehensive body of research detailed in this thesis has successfully and systematically addressed the guiding research questions initially posed in Chapter 1. I have conclusively shown that pre-stimulus EEG indeed harbors significant predictive power

for driver reaction time. This predictive capability was progressively enhanced, starting from classical machine learning, advancing with specialized 1D-CNNs, and culminating in state-of-the-art performance through the novel application of Vision Transformers to image-transformed EEG spectral data. This progression itself forms a key narrative of methodological advancement within the thesis.

For the critical task of drowsiness detection, I empirically confirmed the strong standalone utility of vision data and meticulously demonstrated that while simpler multimodal fusion approaches offer limited incremental benefits, a sophisticated end-to-end multimodal transformer architecture can effectively synergize EEG and vision data to achieve exceptionally high classification accuracy. Finally, bridging the gap to practical application, I successfully engineered and validated an efficient hybrid vision transformer-LSTM model capable of high-accuracy, real-time drowsiness detection on a common mobile edge device.

The overarching contributions of this thesis to the field of intelligent driver state monitoring can be summarized as:

1. A systematic and deep analysis of EEG-based reaction time prediction from pre-stimulus neural signals, including thorough parameter optimization and comparative model evaluations, leading to new insights into EEG feature representation.

2. The novel and successful application of advanced Vision Transformer (ViT-B/16) architectures to image-based representations of EEG spectral features, establishing a new state-of-the-art for EEG-based RT prediction within this work.

3. The development and rigorous validation of a cutting-edge end-to-end multimodal (EEG-Vision) transformer architecture, demonstrating superior performance for robust driver drowsiness detection through effective synergistic fusion.

4. The proposal, implementation, and validation of a novel, efficient hybrid deep learning model (MobileViT-LSTM) specifically designed for practical, real-time, edge-based vision drowsiness detection, complete with deployment considerations like ONNX export.

These contributions, collectively, represent a significant advancement in the application of artificial intelligence to enhance driver safety, covering a spectrum of techniques from foundational neurophysiological signal analysis to the engineering of deployable edge AI systems. The consistent emphasis on robust, subject-independent validation further strengthens the relevance and potential impact of these findings.

8.3 Limitations of the Research

While this thesis presents several impactful findings and contributions, it is essential to acknowledge its inherent limitations, which also serve to illuminate avenues for future scholarly inquiry:

- **Dependence on Simulated Driving Environments:** All empirical studies presented were conducted using data collected within controlled driving simulator environments. While simulators offer distinct advantages for inducing specific driver states (e.g., fatigue) and collecting high-quality, synchronized data, they do not perfectly replicate the full spectrum of cognitive demands, environmental variabilities (e.g., unpredictable lighting, diverse weather, complex road conditions), or the rich array of sensory stimuli encountered in real-world, on-road driving. Consequently, the direct generalizability of the developed models to unconstrained on-road conditions requires explicit and extensive future validation.

- **Dataset Specificity and Diversity of Participant Demographics:** The findings are intrinsically linked to the specific characteristics of the datasets employed (the Cao et al. dataset for RT prediction and the Tobii dataset for drowsiness classification). Although the Tobii dataset featured a commendable age range, both datasets were collected within particular geographical and cultural contexts. The robustness and performance of the proposed models across more diverse driver populations (e.g., varying in driving experience, cultural backgrounds, or health conditions) warrant further investigation through broader data collection efforts.

- **Scope of Modalities and Features Investigated:** In the EEG analyses, the

primary focus was on spectral power features. While these are well-established and informative, other EEG characteristics—such as inter-channel connectivity measures (e.g., coherence, phase-locking value), analyses of microstates, or event-related potentials (ERPs, if the experimental paradigm allowed for more distinctly time-locked cognitive events beyond diffuse state changes)—might offer complementary or even superior predictive information. Similarly, for multimodal fusion, only EEG and facial vision data were considered. The incorporation of additional, readily available modalities (e.g., Electrooculography (EOG) for precise eye movement tracking, physiological signals from consumer wearables like heart rate variability, or even vehicle telemetry data like steering wheel movements and lane position) could potentially further enhance the robustness and accuracy of driver state assessment systems.

- **Computational Demands of Advanced Deep Learning Models:** While Chapter 7 successfully addressed the efficiency challenge for a vision-only edge model, the training of the large-scale end-to-end multimodal transformer detailed in Chapter 6, as well as the fine-tuning of Vision Transformers in Chapter 5, remains computationally intensive, necessitating significant GPU resources and considerable training time. These computational demands can be a barrier for research groups with limited resources and for rapid prototyping.

- **Interpretability of Complex Deep Learning Architectures:** Deep learning models, particularly advanced architectures like Vision Transformers and multimodal transformers, while achieving high predictive accuracy, often operate as "black boxes." Understanding precisely which features the model is learning and how it arrives at its predictions can be challenging. Although outside the primary scope of this thesis, further exploration of model interpretability techniques (e.g., attention map visualization, layer-wise relevance propagation) could yield deeper neurophysiological or behavioural insights and increase trust in these complex systems.

- **Nature of Drowsiness Labeling:** The drowsiness classification in Chapter 6 relied on a binary distinction ('Alert' vs. 'Drowsy') primarily defined by the experimental session times (10 AM vs. 3 AM), albeit validated by KSS scores. Real-world driver drowsiness is a dynamic and continuous spectrum. Future work could aim to develop models capable of predicting finer-grained drowsiness levels or regressing directly onto continuous subjective scales like the KSS.

8.4 Future Research Directions

The findings and identified limitations of this thesis naturally give rise to several compelling and promising avenues for future research, aimed at further advancing the field of intelligent driver state monitoring:

1. **Rigorous On-Road Validation and Deployment Studies:** The paramount next step is the comprehensive validation of the most promising models developed herein—particularly the end-to-end multimodal transformer for drowsiness and the MobileViT-LSTM edge model—using extensive data collected from real-world, on-road driving. Such studies should encompass diverse driving conditions, varied driver demographics, and naturalistic occurrences of fatigue and drowsiness.

2. **Exploration of Advanced Time-Series and Graph-Based EEG Models:** For EEG analysis, future investigations could explore cutting-edge time-series modeling techniques, such as advanced temporal convolutional networks (TCNs) or specialized EEG-centric transformers applied directly to raw or minimally processed EEG epochs, potentially capturing richer temporal dynamics than spectral features alone. Furthermore, Graph Neural Networks (GNNs) offer a powerful framework for explicitly modeling the spatial relationships and dynamic connectivity between EEG channels.

3. **Enhancement of Multimodal Fusion Architectures:** The success of the end-to-end multimodal transformer encourages further research into even more sophisticated fusion mechanisms. This could include exploring adaptive fusion strategies

that dynamically weigh the contribution of each modality based on real-time estimates of signal quality or context, investigating hierarchical fusion approaches, or incorporating attention mechanisms that span longer temporal contexts across modalities. The integration of additional, easily accessible sensor data (e.g., audio cues like speech patterns, physiological data from smartwatches) should also be explored.

4. **Development of Personalized and Adaptive Monitoring Systems:** Addressing the persistent challenge of inter-subject variability is crucial. Future research should focus on developing models that can personalize or adapt to individual drivers over time. Techniques such as transfer learning from general models to specific users with minimal calibration data, few-shot learning, or federated learning approaches (allowing on-device model adaptation while preserving data privacy) hold considerable promise.

5. **Advanced Optimization for Edge AI and Low-Power Systems:** For deployable systems, continued research into advanced model compression techniques beyond ONNX export—such as aggressive quantization (e.g., INT8 or sub-8-bit), network pruning, neural architecture search (NAS) specifically for edge hardware co-design, and knowledge distillation from larger models to compact student models—is essential for minimizing latency and power consumption on embedded automotive platforms or low-power wearable devices.

6. **Prediction of Continuous Drowsiness Levels and Proactive Intervention Strategies:** Moving beyond binary classification, future models should aim to predict continuous levels of drowsiness or fatigue (e.g., by regressing KSS scores or other psychometric scales). This finer-grained assessment can enable more nuanced and timely interventions. Research is also needed on how best to integrate the outputs of these advanced monitoring systems into ADAS to trigger effective and non-distracting alerts, warnings, or even automated vehicle interventions, coupled with studies on optimal Human-Machine Interface (HMI) design for such feedback.

7. **Longitudinal Studies and Understanding Drowsiness Evolution:** Conducting longitudinal studies to track the evolution of drowsiness markers over extended periods (e.g., across an entire work shift for commercial drivers, or during long-haul drives) could provide invaluable data for developing models that understand and predict longer-term fatigue accumulation and its impact on driving safety.

In its entirety, this thesis has endeavored to contribute meaningfully to the dynamic and critically important field of intelligent driver state monitoring. By systematically exploring and advancing a range of techniques, from the fundamental analysis of EEG signals for reaction time prediction to the development of sophisticated multimodal systems for drowsiness detection and efficient edge-AI solutions, this work has aimed to provide a robust foundation for future innovations. It is hoped that the methodologies, findings, and insights presented herein will inspire continued research and development efforts dedicated to leveraging artificial intelligence for the paramount goal of making our roads substantially safer for all users.

Bibliography

[1] World Health Organization. *Road Traffic Injuries*. https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries. Accessed: 2025-06-15. Dec. 2023.

[2] Yong Han Ahn et al. "Factors associated with different levels of daytime sleepiness among Korean construction drivers: a cross-sectional study". In: *BMC public health* 21.1 (2021), pp. 1–12.

[3] Abdulbari Bener et al. "Driver sleepiness, fatigue, careless behavior and risk of motor vehicle crash and injury: Population based case and control study". In: *Journal of Traffic and Transportation engineering (English edition)* 4.5 (2017), pp. 496–502.

[4] AAA Foundation for Traffic Safety. "2017 Traffic Safety Culture Index". In: Accessed: 25/03/2025. Feb. 2018. URL: https://aaafoundation.org/2017-traffic-safety-culture-index/.

[5] Gianluca Borghini et al. "Measuring neurophysiological signals in aircraft pilots and car drivers for the assessment of mental workload, fatigue and drowsiness". In: *Neuroscience Biobehavioral Reviews* 44 (2014), pp. 58–75.

[6] A. M. Williamson and A. M. Feyer. "Moderate sleep deprivation produces impairments in cognitive and motor performance equivalent to legally prescribed levels of alcohol intoxication". In: *Occupational and Environmental Medicine* 57.10 (2000), pp. 649–655.

[7] Anna JHM Beurskens et al. “Fatigue among working people: validity of a questionnaire measure”. In: *Occupational and environmental medicine* 57.5 (2000), pp. 353–357.

[8] Mathias Basner and Joshua Rubinstein. “Fitness for duty: A 3 minute version of the Psychomotor Vigilance Test predicts fatigue related declines in luggage screening performance”. In: *Journal of occupational and environmental medicine/American College of Occupational and Environmental Medicine* 53.10 (2011), p. 1146.

[9] European Parliament and the Council of the European Union. *Regulation (EU) 2019/2144 on type-approval requirements for motor vehicles and their trailers, and systems, components and separate technical units intended for such vehicles, as regards their general safety and the protection of vehicle occupants and vulnerable road users*. Official Journal of the European Union. OJ L 325, 16.12.2019, p. 1–40. Dec. 2019. URL: https://eur-lex.europa.eu/eli/reg/2019/2144/oj.

[10] European Commission. *Commission Delegated Regulation (EU) 2021/1341 of 23 April 2021 supplementing Regulation (EU) 2019/2144 by laying down detailed rules concerning the specific test procedures and technical requirements for the type-approval of motor vehicles with regard to their driver drowsiness and attention warning systems and amending Annex II to that Regulation*. Official Journal of the European Union. OJ L 292, 16.08.2021, p. 4–19. Aug. 2021. URL: https://eur-lex.europa.eu/eli/reg_del/2021/1341/oj.

[11] European New Car Assessment Programme (Euro NCAP). *Assessment Protocol – Safety Assist: Safe Driving, Version 10.3*. Dec. 2023. URL: https://www.euroncap.com/media/79883/euro-ncap-assessment-protocol-sa-safe-driving-v103.pdf (visited on 05/16/2024).

[12] T. Åkerstedt and M. Gillberg. “Subjective and objective sleepiness in the active individual”. In: *International Journal of Neuroscience* 52.1-2 (1990), pp. 29–37.

[13] S. Makeig et al. “Independent component analysis of electroencephalographic data”. In: (1996), pp. 145–151.

[14] Wolfgang Klimesch. “EEG alpha and theta oscillations reflect cognitive and memory performance: a review and analysis”. In: *Brain research reviews* 29.2-3 (1999), pp. 169–195.

[15] C. Cajochen et al. “Power density in theta/alpha frequencies of the waking EEG progressively increases during sustained wakefulness”. In: *Sleep* 18.10 (1995), pp. 890–894.

[16] Q. Ji, Z. Zhu, and P. Lan. “Real-time nonintrusive monitoring and prediction of driver fatigue”. In: *IEEE Transactions on Vehicular Technology* 53.4 (2004), pp. 1052–1068.

[17] P Khunpisuth et al. “Driver drowsiness detection using eye-closing ratio”. In: *2016 12th International Conference on Signal-Image Technology Internet-Based Systems (SITIS)*. IEEE. 2016, pp. 661–668.

[18] Z. Lian et al. “Driving fatigue detection based on hybrid electroencephalography and eye tracking”. In: *IEEE Journal of Biomedical and Health Informatics* 28.1 (2024), pp. 2431–2442.

[19] W. L. Zheng and B. L. Lu. “A multimodal approach to estimating vigilance using EEG and forehead EOG”. In: *Journal of Neural Engineering* 14.2 (2017), p. 026017.

[20] Zehong Cao et al. “Multi-channel EEG recordings during a sustained-attention driving task”. In: *Scientific data* 6.1 (2019), p. 19.

[21] Vivian WY Tam and Ivan WH Fung. “Tower crane safety in the construction industry: A Hong Kong study”. In: *Safety science* 49.2 (2011), pp. 208–215.

[22] Alistair W MacLean, David RT Davies, and Kris Thiele. “The hazards and prevention of driving while sleepy”. In: *Sleep medicine reviews* 7.6 (2003), pp. 507–521.

[23] D. Dawson and K. Reid. “Fatigue, alcohol and performance impairment”. In: *Nature* 388.6639 (1997), p. 235.

[24] J Stephen Higgins et al. “Asleep at the wheel—the road to addressing drowsy driving”. In: *Sleep* 40.2 (2017), zsx001.

[25] National Highway Traffic Safety Administration (NHTSA). “Drowsy Driving”. In: *https://www.nhtsa.gov/risky-driving/drowsy-driving* (n.d.).

[26] Yvonne Tran et al. “The relationship between spectral changes in heart rate variability and fatigue”. In: *Journal of Psychophysiology* 23.3 (2009), pp. 143–151.

[27] J. Vicente et al. “Heart rate variability-based driver drowsiness detection and its validation with EEG”. In: *IEEE Transactions on Biomedical Engineering* 61.3 (2016), pp. 805–811.

[28] S. B. Borgheai et al. “Detecting driver fatigue using heart rate variability: A systematic review”. In: *Accident Analysis & Prevention* 165 (2022), p. 106507.

[29] J. Vicente et al. “Drowsiness detection using heart rate variability”. In: *Medical & Biological Engineering & Computing* 54.6 (2016), pp. 927–937.

[30] Yair Morad et al. “Pupillography as an objective indicator of fatigue”. In: *Current eye research* 21.1 (2000), pp. 535–542.

[31] S. F. Hendi and B. Y. Majlis. “Development of vehicle driver drowsiness detection system using electrooculogram (EOG)”. In: *Proceedings of the 2005 1st International Conference on Computers, Communications, & Signal Processing with Special Track on Biomedical Engineering.* 2005, pp. 165–168.

[32] Y. Tian and J. Cao. “Fatigue driving detection based on electrooculography: a review”. In: *EURASIP Journal on Image and Video Processing* 2021.1 (2021), p. 33.

[33] Sadegh Arefnezhad et al. “Driver drowsiness detection based on steering wheel data applying adaptive neuro-fuzzy feature selection”. In: *Sensors* 19.4 (2019), p. 943.

[34] Chun-Shu Wei et al. “A subject-transfer framework for obviating inter-and intra-subject variability in EEG-based drowsiness detection”. In: *NeuroImage* 174 (2018), pp. 407–419.

[35] Chun-Shu Wei et al. “Toward drowsiness detection using non-hair-bearing EEG-based brain-computer interfaces”. In: *IEEE transactions on neural systems and rehabilitation engineering* 26.2 (2018), pp. 400–406.

[36] Izzat A Akbar et al. “Three drowsiness categories assessment by electroencephalogram in driving simulator environment”. In: *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE. 2017, pp. 2904–2907.

[37] Shuyan Hu and Gangtie Zheng. “Driver drowsiness detection with eyelid related parameters by Support Vector Machine”. In: *Expert Systems with Applications* 36.4 (2009), pp. 7651–7658.

[38] Saeed Abtahi et al. “YAWDD: A yawning detection dataset”. In: *Proceedings of the 5th ACM Multimedia Systems Conference*. 2014, pp. 24–28.

[39] F. Wang, L. Zheng, and Y. Liu. “Driver fatigue detection based on facial features and driving performance”. In: *International Journal of Digital Content Technology and its Applications* 10.5 (2016), pp. 1–10.

[40] K. Dwivedi, K. Biswaranjan, and A. Sethi. “Drowsy driver detection using representation learning”. In: (2014), pp. 995–999.

[41] S. Abtahi, B. Hariri, and S. Shirmohammadi. “Driver drowsiness monitoring based on yawning detection”. In: *2011 IEEE International Instrumentation and Measurement Technology Conference*. 2011, pp. 1–4.

[42] A. Muhammad, N. Badruddin, and M. Drieberg. “Driver drowsiness detection using EEG power spectrum analysis”. In: (2014), pp. 300–303.

[43] M. Zhu et al. “Vehicle driver drowsiness detection method using wearable EEG based on convolution neural network”. In: *Neural Computing and Applications* 33 (2021), pp. 13965–13980.

[44] Q. Rezaee et al. *Driver drowsiness detection with commercial EEG headsets*. arXiv preprint arXiv:2303.14841. 2023.

[45] Kuan-Chih Huang et al. “The effects of different fatigue levels on brain–behavior relationships in driving”. In: *Brain and Behavior* 9.12 (2019), e01379. ISSN: 2162-3279. DOI: 10.1002/brb3.1379.

[46] Michael A Vidulich et al. “Performance-based and physiological measures of situational awareness.” In: *Aviation, space, and environmental medicine* (1994).

[47] Torbjöm Åkerstedt, Göran Kecklund, and Anders Knutsson. “Manifest sleepiness and the spectral content of the EEG during shift work”. In: *Sleep* 14.3 (1991), pp. 221–225.

[48] Y. Tran et al. “The influence of mental fatigue on brain activity: Evidence from a systematic review with meta-analyses”. In: *Psychophysiology* 57.5 (2020).

[49] Mike X Cohen. *Analyzing neural time series data: theory and practice*. MIT press, 2014.

[50] A. Bashashati et al. “A survey of signal processing algorithms in brain–computer interfaces based on electrical brain signals”. In: *Journal of Neural Engineering* 4.2 (2007), R32–R57.

[51] Jian Cui et al. “A compact and interpretable convolutional neural network for cross-subject driver drowsiness detection from single-channel EEG”. In: *Methods* (2021).

[52] Ruyi Foong, Kai Keng Ang, and Chai Quek. “Correlation of reaction time and EEG log bandpower from dry frontal electrodes in a passive fatigue driving simulation experiment”. In: *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE. 2017, pp. 2482–2485.

[53] Yisi Liu et al. “Inter-subject transfer learning for EEG-based mental fatigue recognition”. In: *Advanced Engineering Informatics* 46 (2020), p. 101157.

[54] Yisi Liu et al. “EEG-based cross-subject mental fatigue recognition”. In: *2019 International Conference on Cyberworlds (CW)*. IEEE. 2019, pp. 247–252.

[55] Rafał S Jurecki and Tomasz L Stańczyk. "Driver reaction time to lateral entering pedestrian in a simulated crash traffic situation". In: *Transportation research part F: traffic psychology and behaviour* 27 (2014), pp. 22–36.

[56] Alexandre Gramfort et al. "MNE software for processing MEG and EEG data". In: *Neuroimage* 86 (2014), pp. 446–460.

[57] O. M. Solomon Jr. *PSD computations using Welch's method*. Tech. rep. 1991. DOI: 10.2172/5688766.

[58] Y. Roy et al. "Deep learning-based electroencephalography analysis: A systematic review". In: *Journal of Neural Engineering* 16.5 (2019), p. 051001.

[59] A. Craik, Y. He, and J. L. Contreras-Vidal. "Deep learning for electroencephalogram (EEG) classification tasks: a review". In: *J. Neural Eng.* 16.3 (2019), p. 031001.

[60] Vernon J Lawhern et al. "EEGNet: a compact convolutional neural network for EEG-based brain–computer interfaces". In: *Journal of neural engineering* 15.5 (2018), p. 056013.

[61] Jian Cui et al. "EEG-based Cross-Subject Driver Drowsiness Recognition with Interpretable CNN". In: *arXiv preprint arXiv:2107.09507* (2021).

[62] Yanina Atum et al. "A comparison of subject-dependent and subject-independent channel selection strategies for single-trial p300 brain computer interfaces". In: *Medical & biological engineering & computing* 57 (2019), pp. 2705–2715.

[63] V. Jayaram et al. "Transfer learning in brain-computer interfaces". In: *IEEE Computational Intelligence Magazine* 11.1 (2016), pp. 20–31.

[64] F. Lotte et al. "A review of classification algorithms for EEG-based brain-computer interfaces". In: *Journal of Neural Engineering* 4.2 (2007), R1–R13.

[65] Wolfgang Klimesch. "Alpha-band oscillations, attention, and controlled access to stored information". In: *Trends in cognitive sciences* 16.12 (2012), pp. 606–617.

[66] Geoffrey F Woodman. “A brief introduction to the use of event-related potentials in studies of perception and attention”. In: *Attention, Perception, & Psychophysics* 72 (2010), pp. 2031–2046.

[67] Alan Gevins and Michael E Smith. “Neurophysiological measures of cognitive workload during human-computer interaction”. In: *Theoretical issues in ergonomics science* 4.1-2 (2003), pp. 113–131.

[68] Raul Fernandez Rojas et al. “Electroencephalographic workload indicators during teleoperation of an unmanned aerial vehicle shepherding a swarm of unmanned ground vehicles in contested environments”. In: *Frontiers in neuroscience* 14 (2020), p. 40.

[69] Stefanie Enriquez-Geppert, Rene J Huster, and Christoph S Herrmann. “EEG-neurofeedback as a tool to modulate cognition and behavior: A review tutorial”. In: *Frontiers in human neuroscience* 11 (2017), p. 51.

[70] R. J. Barry, A. R. Clarke, and S. J. Johnstone. “A review of electrophysiology in attention-deficit/hyperactivity disorder: II. Event-related potentials”. In: *Clinical neurophysiology* 114.2 (2003), pp. 184–198.

[71] Matthew D Sacchet et al. “Attention drives synchronization of alpha and beta rhythms between right inferior frontal and primary sensory neocortex”. In: *Journal of neuroscience* 35.5 (2015), pp. 2074–2082.

[72] Markus Bauer et al. “Attentional modulation of alpha/beta and gamma oscillations reflect functionally distinct processes”. In: *Journal of Neuroscience* 34.48 (2014), pp. 16117–16125.

[73] Elisa Magosso, Giulia Ricci, and Mauro Ursino. “Alpha and theta mechanisms operating in internal-external attention competition”. In: *Journal of Integrative Neuroscience* 20.1 (2021), pp. 1–19.

[74] Zeynep A Acar and Scott Makeig. “Effects of forward model errors on EEG source localization”. In: *Brain topography* 23.1 (2010), pp. 73–86.

[75] Sheng-Fu Liang, Hsu-Chuan Wang, and Wan-Lin Chang. "Combination of EEG complexity and spectral analysis for epilepsy diagnosis and seizure detection". In: *EURASIP journal on advances in signal processing* 2010 (2010), pp. 1–15.

[76] Richard W Homan, John Herman, and Phillip Purdy. "Cerebral location of international 10–20 system electrode placement". In: *Electroencephalography and clinical neurophysiology* 66.4 (1987), pp. 376–382.

[77] Abhijit Rajan et al. "Theta oscillations index frontal decision-making and mediate reciprocal frontal–parietal interactions in willed attention". In: *Cerebral Cortex* 29.7 (2019), pp. 2832–2843.

[78] Wang Wan et al. "Frontal EEG-based multi-level attention states recognition using dynamical complexity and extreme gradient boosting". In: *Frontiers in Human Neuroscience* 15 (2021), p. 673955.

[79] Hannah R Sheahan et al. "Imagery of movements immediately following performance allows learning of motor skills that interfere". In: *Scientific Reports* 8.1 (2018), pp. 1–12.

[80] Xiangbin Teng et al. "Concurrent temporal channels for auditory processing: Oscillatory neural entrainment reveals segregation of function at different scales". In: *PLoS biology* 15.11 (2017), e2000812.

[81] Carole Peyrin and Benoit Musel. "On the specific role of the occipital cortex in scene perception". In: *Visual cortex: current status and perspectives (Molotchnikoff S, Rouat J, eds)* (2012), pp. 61–82.

[82] Alberto Zani and Alice Mado Proverbio. "Spatial attention modulates earliest visual processing: An electrical neuroimaging study". In: *Heliyon* 6.11 (2020), e05570.

[83] Gahangir Hossain, Mark H Myers, and Robert Kozma. "Spatial directionality found in frontal-parietal attentional networks". In: *Neuroscience journal* 2018 (2018).

[84] Maarten Schrooten et al. “Electrocorticography of spatial shifting and attentional selection in human superior parietal cortex”. In: *Frontiers in Human Neuroscience* 11 (2017), p. 240.

[85] Maarten De Vos and Stefan Debener. “Towards a truly mobile auditory brain-computer interface: exploring the P300 to take away spatial restrictions”. In: *PLoS One* 9.8 (2014), e105055.

[86] Zahra Mardi, Seyedeh Naghmeh Miri Ashtiani, Mohammad Mikaili, et al. “EEG-based drowsiness detection for safe driving using chaotic features and statistical tests”. In: *Journal of Medical Signals & Sensors* 1.2 (2011), p. 130.

[87] Benjamin Blankertz et al. “Optimizing spatial filters for robust EEG single-trial analysis”. In: *IEEE Signal Processing Magazine* 25.1 (2008), pp. 41–56.

[88] Andrey Andreev et al. “Common spatial patterns revisited: consistency and stability”. In: *Journal of Neural Engineering* 16.2 (2019), p. 026026.

[89] Wanze Xie, Russell T Toll, and Charles A Nelson. “EEG functional connectivity analysis in the source space”. In: *Developmental Cognitive Neuroscience* 56 (2022), p. 101119.

[90] Leonardo Rundo and Carmelo Militello. “Image Biomarkers and Explainable AI: Handcrafted Features versus Deep Learned Features”. In: *European Radiology Experimental* 8.1 (2024), pp. 1–10. DOI: 10.1186/s41747-024-00529-y.

[91] Tereza Soukupova and Jan Cech. “Eye blink detection using facial landmarks”. In: *21st computer vision winter workshop, Rimske Toplice, Slovenia*. Vol. 2. 2016, p. 4.

[92] Biying Fu et al. “A survey on drowsiness detection–modern applications and methods”. In: *IEEE Transactions on Intelligent Vehicles* (2024).

[93] Xu-Yao Zhang, Cheng-Lin Liu, and Ching Y. Suen. “Towards Robust Pattern Recognition: A Review”. In: *arXiv preprint arXiv:2006.06976* (2020). URL: https://arxiv.org/abs/2006.06976.

[94] Rishabh Kumar et al. “Driver drowsiness detection system using image processing”. In: *2023 12th International Conference on System Modeling & Advancement in Research Trends (SMART)*. IEEE. 2023, pp. 19–22.

[95] Muneeb Ahmed et al. “Intelligent driver drowsiness detection for traffic safety based on multi CNN deep model and facial subsampling”. In: *IEEE transactions on intelligent transportation systems* 23.10 (2021), pp. 19743–19752.

[96] Ken Alparslan, Yigit Alparslan, and Matthew Burlick. “Towards evaluating driver fatigue with robust deep learning models”. In: *arXiv preprint arXiv:2007.08453* (2020).

[97] Kaiming He et al. “Deep residual learning for image recognition”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016, pp. 770–778.

[98] S. Ji et al. “3D convolutional neural networks for human action recognition”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35.1 (2013), pp. 221–231.

[99] A. Karpathy et al. “Large-scale video classification with convolutional neural networks”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2014, pp. 1725–1732.

[100] B. Baheti, S. Gajre, and S. Talbar. “Detection of distracted driver using convolutional neural network”. In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 2018, pp. 1145–1151.

[101] Alexey Dosovitskiy et al. “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale”. In: *International Conference on Learning Representations (ICLR)*. 2021. URL: https://arxiv.org/abs/2010.11929.

[102] Salman Khan et al. “Transformers in vision: A survey”. In: *ACM computing surveys (CSUR)* 54.10s (2022), pp. 1–41.

[103] Seung Il Lee et al. “Vision transformer models for mobile/edge devices: a survey”. In: *Multimedia Systems* 30.2 (2024), p. 109.

[104] Wenchao Xu et al. “Deploying Foundation Model Powered Agent Services: A Survey”. In: *arXiv preprint arXiv:2412.13437* (2024).

[105] R. J. Deligani et al. “Multimodal fusion of EEG-fNIRS: a mutual information-based hybrid classification framework”. In: *Biomedical Optics Express* 12.3 (2021), pp. 1635–1650.

[106] C. T. Lin et al. “Adaptive EEG-based alertness estimation system by using ICA-based fuzzy neural networks”. In: *IEEE Transactions on Circuits and Systems I: Regular Papers* 53.11 (2005), pp. 2469–2476.

[107] Y. P. Lin and T. P. Jung. “Improving EEG-based emotion classification using conditional transfer learning”. In: *Frontiers in Human Neuroscience* 11 (2017), p. 334.

[108] X. Zhang et al. “Driver drowsiness detection using multi-channel second-order blind identification and convolutional neural networks”. In: *Expert Systems with Applications* 169 (2021), p. 114283.

[109] M. E. Tipping. “Sparse Bayesian learning and the relevance vector machine”. In: *Journal of Machine Learning Research* 1 (June 2001), pp. 211–244.

[110] A. Vaswani, N. Shazeer, N. Parmar, et al. “Attention is all you need”. In: *Advances in Neural Information Processing Systems*. 2017, pp. 5998–6008.

[111] A. Dosovitskiy, L. Beyer, A. Kolesnikov, et al. *An image is worth 16x16 words: Transformers for image recognition at scale*. arXiv preprint arXiv:2010.11929. 2020.

[112] Y. H. H. Tsai et al. “Multimodal transformer for unaligned multimodal language sequences”. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 2019, pp. 6558–6569.

[113] T. Song et al. “EEG emotion recognition using dynamical graph convolutional neural networks”. In: *IEEE Transactions on Affective Computing* 11.3 (2018), pp. 532–541.

[114] Sachin Mehta et al. "MobileViT: Light-weight, General-purpose, and Mobile-friendly Vision Transformer". In: *arXiv preprint arXiv:2110.02178* (2021). URL: https://arxiv.org/abs/2110.02178.

[115] Reza Ghoddoosian, Marnim Galib, and Vassilis Athitsos. "A realistic dataset and baseline temporal model for early drowsiness detection". In: *Proceedings of the ieee/cvf conference on computer vision and pattern recognition workshops*. 2019, pp. 0–0.

[116] Yini Deng, Yingying Jiao, and Bao-Liang Lu. "Driver sleepiness detection using lstm neural network". In: *Neural Information Processing: 25th International Conference, ICONIP 2018, Siem Reap, Cambodia, December 13-16, 2018, Proceedings, Part IV 25*. Springer. 2018, pp. 622–633.

[117] Sepp Hochreiter and Jürgen Schmidhuber. "Long Short-Term Memory". In: *Neural Computation* 9.8 (1997), pp. 1735–1780. DOI: 10.1162/neco.1997.9.8.1735.

[118] Ming-Zhou Liu et al. "Real time detection of driver fatigue based on CNN-LSTM". In: *IET Image Processing* 16.2 (2022), pp. 576–595.

[119] Mohamed Waheed Gomaa, Rasha O Mahmoud, and Amany M Sarhan. "A cnn-lstm-based deep learning approach for driver drowsiness prediction". In: *Journal of Engineering Research* 6.3 (2022), pp. 59–70.

[120] Yin-Cheng Tsai et al. "Vision-based instant measurement system for driver fatigue monitoring". In: *IEEE Access* 8 (2020), pp. 67342–67353.

[121] Paulius Micikevicius et al. "Mixed Precision Training". In: *International Conference on Learning Representations (ICLR)*. 2018. URL: https://arxiv.org/abs/1710.03740.

[122] Benoit Jacob et al. "Quantization and training of neural networks for efficient integer-arithmetic-only inference". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 2704–2713.

[123] Jiaxiang Wu et al. “Quantized convolutional neural networks for mobile devices”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 4820–4828.

[124] *ONNX: Open Neural Network Exchange*. `https://onnx.ai/`. Accessed: 25/03/2025. 2017.

[125] Anastasios Fanariotis et al. “Power efficient machine learning models deployment on edge IoT devices”. In: *Sensors* 23.3 (2023), p. 1595.

[126] Xiao Jiang, Gui-Bin Bian, and Zean Tian. “Removal of artifacts from EEG signals: a review”. In: *Sensors* 19.5 (2019), p. 987.

[127] Morteza Zangeneh Soroush et al. “EEG artifact removal using sub-space decomposition, nonlinear dynamics, stationary wavelet transform and machine learning algorithms”. In: *Frontiers in Physiology* 13 (2022), p. 910368.

[128] Nurhan Gursel Ozmen, Levent Gumusel, and Yuan Yang. “A biologically inspired approach to frequency domain feature extraction for EEG classification”. In: *Computational and mathematical methods in medicine* 2018 (2018).

[129] M. S. N. Chowdhury. “Deep Neural Network for Visual Stimulus-Based Reaction Time Estimation Using the Periodogram of Single-Trial EEG”. In: *IEEE Transactions on Biomedical Engineering*. Vol. 67. 11. 2020, pp. 3288–3297.

[130] J. J. Foxe and A. C. Snyder. “The role of alpha-band brain oscillations as a sensory suppression mechanism during selective attention”. In: *Frontiers in Psychology* 2 (2011), p. 154.

[131] R. T. Schirrmeister, J. T. Springenberg, L. Dieleman, et al. “Deep learning with convolutional neural networks for EEG decoding and visualization”. In: *Hum. Brain Mapp.* 38.11 (2017), pp. 5391–5420.

[132] J. Deng et al. “ImageNet: A large-scale hierarchical image database”. In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 2009, pp. 248–255.

[133] D. P. Kingma and J. Ba. *Adam: A method for stochastic optimization*. arXiv preprint arXiv:1412.6980. 2014.

[134] X. Glorot and Y. Bengio. "Understanding the difficulty of training deep feedforward neural networks". In: *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*. 2010, pp. 249–256.

[135] A. Paszke, S. Gross, F. Massa, et al. "PyTorch: An imperative style, high-performance deep learning library". In: *Advances in Neural Information Processing Systems*. 2019, pp. 8024–8035.

[136] Z. Lv et al. "Investigating critical brain area for EEG-based binocular color fusion and rivalry with EEGNet". In: *Frontiers in Neuroscience* 18 (2024), p. 1361486.

[137] A.-C. Phan et al. "An efficient approach for detecting driver drowsiness based on deep learning". In: *Applied Sciences* 11.18 (2021), p. 8441.

[138] Xiaoyu Ren et al. "FCN+: Global receptive convolution makes fcn great again". In: *Neurocomputing* 631 (2025), p. 129655.

[139] Agustina Garcés Correa, Lorena Orosco, and Eric Laciar. "Automatic detection of drowsiness in EEG records based on multimodal analysis". In: *Medical Engineering Physics* 36.2 (Feb. 2014), pp. 244–249. ISSN: 1350-4533. DOI: 10.1016/j.medengphy.2013.07.011. URL: https://www.sciencedirect.com/science/article/pii/S1350453313001690.

[140] Thien Nguyen et al. "Utilization of a combined EEG/NIRS system to predict driver drowsiness". In: *Scientific Reports* 7 (Mar. 2017), p. 43933. ISSN: 2045-2322. DOI: 10.1038/srep43933. URL: https://www.nature.com/articles/srep43933.

[141] S. Sheykhivand et al. "Automatic Detection of Driver Fatigue Based on EEG Signals Using a Developed Deep Neural Network". In: *Electronics* 11.14 (2022), p. 2169. DOI: 10.3390/electronics11142169. URL: https://www.mdpi.com/2079-9292/11/14/2169.

[142] Ilya Loshchilov and Frank Hutter. “Decoupled Weight Decay Regularization”. In: *arXiv preprint arXiv:1711.05101* (2017). Published at ICLR 2019. URL: https://arxiv.org/abs/1711.05101.