Cognitive Load Estimation from Eye-Tracking Data via Cross-Domain & Test-Time Adaptation

by

Shafat Ahnaf Mufdi
22301021
Zarif Ashrafee
22301222
Ahmed Lamha
22301148
Md. Faisal Islam
22301011

A thesis submitted to the
Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science and Engineering

Department of Computer Science and Engineering
BRAC University
January 2026.

© 2026. BRAC University
All rights reserved.

Declaration

It is hereby declared that

1. The thesis submitted is our own original work while completing our degrees at BRAC University.
2. The thesis does not contain material previously published or written by a third party, except where it is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

Student's Full Name & Signature:

Shafat Ahnaf Mufdi
22301021

Zarif Ashrafee
22301222

Ahmed Lamha
22301148

Md. Faisal Islam
22301011

Approval

The thesis titled "Cognitive Load Estimation from Eye-Tracking Data via Cross-Domain & Test-Time Adaptation" submitted by

1. Shafat Ahnaf Mufdi (22301021)
2. Zarif Ashrafee (22301222)
3. Ahmed Lamha (22301148)
4. Md. Faisal Islam (22301011)

has been accepted as satisfactory in partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering in Fall 2025.

Examining Committee:

Supervisor:
(Member)

Md. Sabbir Ahmed

Senior Lecturer
Department of Computer Science and Engineering
BRAC University

Thesis Coordinator:
(Member)

Dr. Md. Golam Rabiul Alam

Professor
Department of Computer Science and Engineering
BRAC University

Head of Department:
(Chair)

Dr. Sadia Hamid Kazi

Associate Professor & Chairperson
Department of Computer Science and Engineering
BRAC University

Abstract

While the estimation of cognitive load through eye-tracker data has proven to hold immense potential for Human-Computer Interaction, existing implementations are often hindered by significant inter-subject variability and the need for calibration. Domain adaptation techniques attempt to mitigate these shifts by aligning source & target distributions globally. However, this creates a static model that is unable to account for instance-specific variations, namely physiological differences, lighting changes, sensor dropouts, and changes in hardware setups. Integrating test-time adaptation into the pipeline allows for real-time normalization of the model's statistics using unlabeled streaming data from the test subject, which has been a challenge in prior studies on creating impartial cognitive load classifiers. This makes the model dynamic, which is validated by benchmarking using the COLET & ADABase datasets, representing varied scenarios, subjects, and setups. Our results show that adding the dimension of test-time adaptation improves generalization accuracy without requiring explicit recalibration, which is crucial as the use of extended reality (XR) headsets and the applications of situationally-aware interfaces are rapidly becoming widespread.

Keywords: cognitive load theory (CLT), eye-tracking, domain adaptation, correlation alignment (CORAL), test-time adaptation (TTA), iterative pseudo-labeling, situationally-induced impairments & disabilities (SIID)

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

Contextual or environmental factors may momentarily disrupt an individual's sensory and cognitive capacities, potentially leading to unsatisfactory experiences and temporary limitations. These phenomena are collectively known as Situationally-induced impairments and disabilities (SIIDs) [36]. Namely, cognitive overload, motion constraints, low light, or even noise can lead to situational impairments. In the rapidly evolving landscape of Human Computer Interaction (HCI), particularly within Extended Reality (XR) or domains where attention is critical, like aviation or driving, the ability to monitor a user's cognitive state is paramount. While frameworks like Human I/O have attempted to detect the physical constraints of availability, they have overlooked the internal and functional state of the user, which is the measure of Cognitive Load (CL). As CL is the total amount of mental effort being used in the working memory, high CL can act as a silent yet predominant impairment and effectively reduce the user's ability to process visual or auditory information, even if their physical channels are available. Recent studies have proven eye-tracking data as a powerful and non-intrusive modality for successfully estimating CL as ocular metrics like pupil dilation and blink rates are involuntary physiological responses to mental load [7]. However, a critical challenge of cross-subject or cross-domain generalization still remains. Eye movements are heavily influenced by the task at hand and the type of task as well [22], it is also affected by environmental lighting, among other external stressors. When classifiers trained in controlled lab environments are deployed in dynamic real-world situations, they often fail due to the shift in domain and to address this, domain adaptation as well as test-time adaptation techniques offer a pathway to create robust, calibration-free CL estimators based on eye-tracking data that can adapt to new users and scenarios in real-time without manual intervention.

1.2 Rationale of the Study

In the rapidly evolving landscape of Human-Computer Interaction (HCI), the accurate estimation of Cognitive Load (CL) has become a critical requirement for maintaining user safety and ensuring accessibility to enhance user experience. While EEG is a proven gold standard for CL detection [6], its intrusiveness limits practical deployment. In contrast, eye-tracking data has been proven to be a sufficient source for CL detection, while being a non-obstructive & ubiquitous alternative [26] suitable for integration into real-world technologies like wearable devices. However, current eye-tracking-based CL estimators suffer from a significant domain shift gap [22], where a model trained in one environment fails to generalize to another. CL estimators also typically require tedious manual interventions to account for individual physiological differences. In dynamic environments where lighting or tasks change rapidly, users can not be expected to pause and recalibrate the system.

Hence, to bridge this gap, our study aims to create a robust & dynamic CL estimator framework. The primary goal of this study is to build a model that learns objective traits of CL that persist across different target spaces by implementing Cross-Domain Adaptation. To make the model self-supervised, the next goal is to integrate Test-Time Adaptation (TTA) to dynamically normalize the statistical distributions. This enables the system to self-calibrate in real-time, using high-confidence predictions, whilst also accounting for temporal shifts. This approach moves beyond simple CL estimation and aims to create a pipeline that ensures reliable inference even in unknown environments & setups, without the need for user interventions.

1.3 Problem Statement

Existing implementations of CL classifiers suffer from inter-subject variances & performance gaps during domain shifts, despite eye-tracking data enabling a viable window into measuring CL. A drastic drop in accuracy is observed when the source model is applied to the target domain due to the change in the statistical distribution of eye features [22], validating the need for adaptation across domains. Prior solutions relied on Z-score standardization methods [22], which necessitates a manual calibration phase to calculate user statistics, and as a result, this renders the system unreliable during the initial phase of interaction. Additionally, static models fail to account for instance-specific variations, leading to misclassifications that can trigger erroneous interface adaptations. As of now, there is a lack of a unified framework that bridges the gap between source & target domains while also handling live fluctuations in estimating CL via eye-tracking data.

1.4 Objective

The objective is to develop a dynamic CL estimation pipeline capable of adapting to different environments without requiring explicit user calibration. Establishing this dimension in CL estimation paves the way for powering the next generation of HCI research on wearable devices with improved & robust context-aware computing in real-world scenarios. The key objectives are as follows:

- Harmonize different eye-tracking datasets into a fused feature space to resolve unit discrepancies.
- Justify the need for adaptation by conducting a Domain Shift Analysis (via External Validation) and statistically quantify the performance degradation of static models when transferred to unseen data.
- Implement Cross-Domain Adaptation to minimize the distance between the covariance matrices of the source and target domains.
- Develop a Test-Time Adaptation mechanism, enabling the model to dynamically adapt to new users' physiological traits in real time.
- Integrate Explainability (XAI) to validate the model's reliance on genuine physiological biomarkers rather than environmental artifacts.

1.5 Methodology in Brief

To estimate cognitive load from raw eye-tracking data without calibration, the pipeline starts with the ingestion of heterogeneous data and proceeds to preprocessing & feature engineering, where important physiological markers are derived from raw eye-tracker data, as shown in Figure 1.1. To make the model independent of subjects or hardware setups, the data is then aligned, where a common feature space is established. The key step in this process flow is the adaptive modeling phase, where, if a domain gap is identified during external validation, the system applies Domain Adaptation to statistically match the source & target distributions. Finally, the system is then adapted at test-time to the live target stream, where the model self-calibrates using high-confidence predictions, for reliable cognitive load classification.

A flowchart of the described brief methodology for CL estimation:

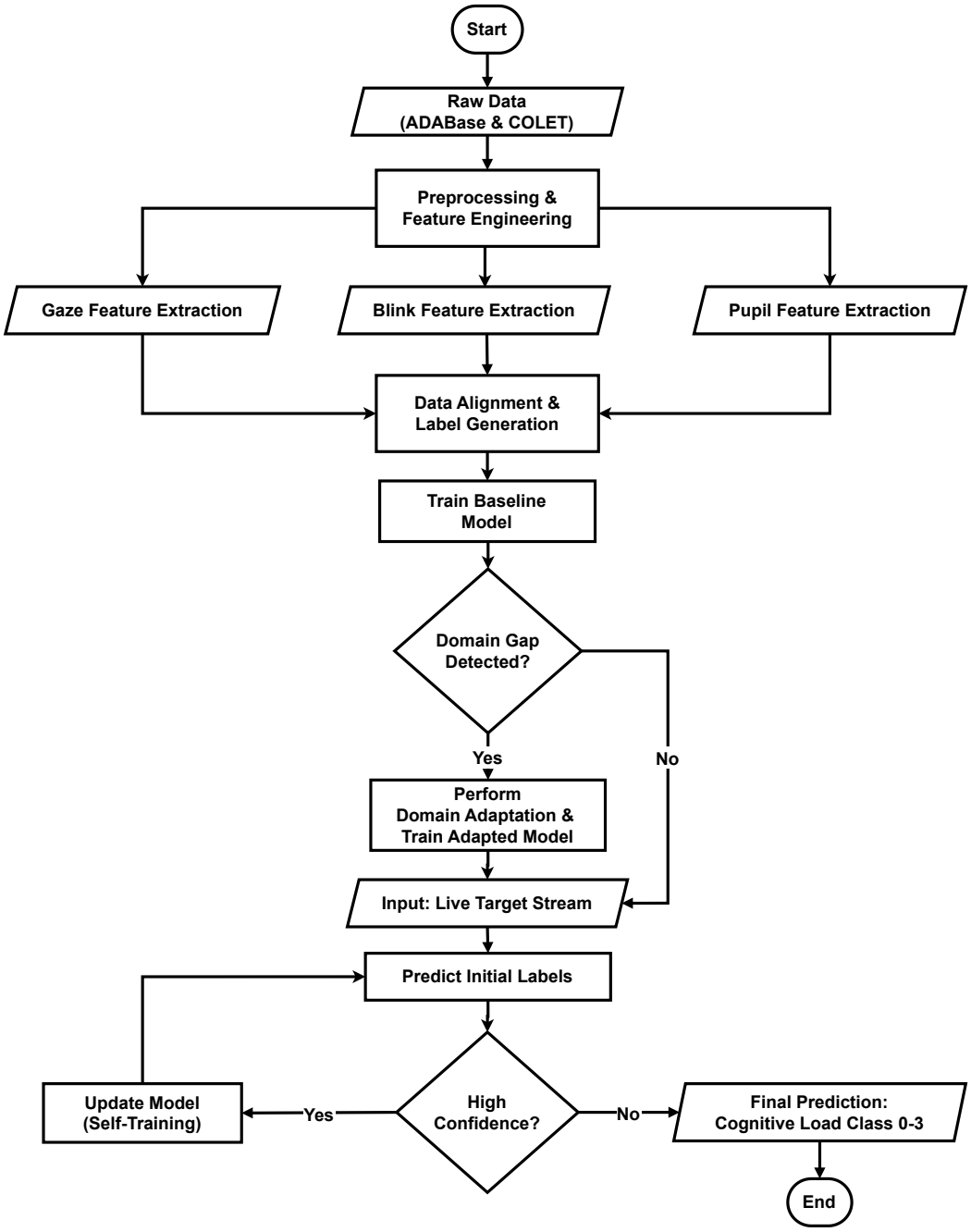

Figure 1.1: CL Estimation Logic Overview

1.6 Scopes and Challenges

This study revolves around the analysis of eye-tracking data for the purpose of multi-class CL classification. The focus is on demonstrating the efficacy of lightweight DA & TTA methods to mitigate the domain shift performance gap, whilst allowing for inference without requiring explicit ground-truth labels. Furthermore, CL estimators exhibiting reliance on domain-dependent features is a challenge, as that results in biased predictions. This study also aims to explore the feature reliance of the proposed models to interpret the effectiveness of the domain adaptation strategies. Another challenge is the alignment of the feature spaces of various eye-tracker datasets, which requires handling in the preprocessing steps of the proposed methodology.

Beyond the scope are - integration of multimodal sensor fusion, analyzing time series EEG data against tabular eye-tracker data for effectiveness in estimating CL, comparing Deep Learning-based Cross-Domain Adaptation & Test-Time Adaptation against lightweight ML-based ensemble models (DANN & TENT vs. CORAL & Iterative Pseudo-Labeling), developing a privacy-preserving framework for eye-tracker-based CL estimation.

Chapter 2

Literature Review

2.1 Preliminaries

Cognitive Load Theory

The quantitative measure of mental fatigue, capacity, and effort required to complete a task is connoted by Cognitive load [24]. Cognitive Load Theory (CLT), introduced in the late 1980s, can be categorized into three types: intrinsic load, which pertains to the complexity of the content; extraneous load, arising from irrelevant external factors; and germane load, which involves the effective construction of schemas through the integration of new and existing information [37]. A fundamental aspect of CLT is its emphasis on working memory, which operates within a limited capacity, and exceeding this threshold adversely affects learning and task performance. Rapid task responses serve as indicators of cognitive stress associated with these limitations. In a rather interesting turn, cognitive load is not congruent across different individuals since it can be influenced by both the individual's prior knowledge and also their personal cognitive ability [37]. Hence, managing cognitive load efficiently, whether it be in professions demanding high levels of engagement and responsibility or in everyday tasks where performance and accessibility may be enhanced, ultimately allows a seamless transition for communication and navigation within an ever-more device and technology-oriented world.

Multimodal Interaction

Multimodal Interaction refers to the process of integrating a quantity of input and output modes, such as speech, gestures, gaze, and haptic feedback, to foster more natural and intuitive Human-Computer Interaction. Uniting voice input with visual or tactile feedback, virtual assistants today, like Amazon Alexa and Google Assistant enable efficiency in interaction by allowing users to interact with systems through different channels simultaneously; thus, it promotes accessibility, adaptability, and the overall user experience. Augmented reality (AR) systems also similarly consolidate gaze and hand gestures for an immersive control experience, as can be seen in Microsoft HoloLens applications [41].

NASA-TLX

The NASA Task Load Index (NASA-TLX) is a widely established multidimensional assessment tool designed to rate the subjective perceived workload of a task by deriving an overall workload score based on a weighted average of six sub-scales comprising Mental Demand, Physical Demand, Temporal Demand, Performance, Effort, and Frustration [3]. In our study, the NASA-TLX score serves as the ground-truth label for the models, since physiological signals like pupil diameter need to be mapped to a verifiable state of CL.

Domain Adaptation

Cross-Domain Adaptation is a Transfer Learning technique, designed to bridge the distribution gap between a labeled source domain and an unlabeled target domain [45]. Unlike standard supervised learning, which assumes training and test data share the same statistical properties, Domain Adaptation addresses the scenario where the target distribution deviates from the source, and it involves accessing a batch of unlabeled data from the target environment during the training phase. Therefore, incorporating these target samples alongside the source data enables the model to align its decision boundaries to the target's specific characteristics before deployment. As a result, it mitigates the risk of performance degradation caused by a shift in input space and ensures the classifier is not biased solely towards the source domain but is effectively adapted to the new user [39].

Correlation Alignment

Correlation Alignment or CORAL, takes the covariance matrices from the subject and the target, finds & applies a linear transformation to minimize differences, and transforms the source domain to align it with the target's statistical structure. Using CORAL, a shared feature space can be created to minimize subject-induced biases and prevent the model from learning traits that may not generalize well across a wider test set [40]. This ensures that when the model is trained on the fused data, it learns features that are statistically consistent, preventing the model from failing due to simple covariate shifts.

Domain-Adversarial Neural Network

Domain-Adversarial Neural Network, or DANN, trains a model to learn domain-invariant features by adding a domain classifier and applying a gradient reversal. During training, the model learns to perform a main task, such as classification. It also learns to trick the domain classifier, which tries to differentiate between the source and the target domains. This adversarial process pushes the model to extract features that are useful for the main classification task but not tied to a specific domain. As a result, performance on new, unseen data is improved, and both subject-induced & domain-induced bias are minimized [44]. As this is a Deep Learning approach, it is best suited for time-series data like EEG signals, where the availability of samples is not an issue [43].

Test-Time Adaptation

While Cross-Domain adaptation techniques help with aligning distributions globally during training, they only produce a static model that can not account for temporal variations. Test-Time Adaptation or TTA, addresses this limitation by updating the pre-trained model continuously during the inference phase [39]. TTA spontaneously fine-tunes the classifier using high-confidence predictions from the incoming data stream, transforming the classifier from a static artifact into a dynamic system capable of automatic self-calibration in real-time for CL estimation and ensuring reliable performance despite any physiological drift.

Iterative Pseudo-Labeling

Iterative Pseudo-Labeling refers to a dynamic, self-training strategy used to adapt a pre-trained classifier to an unlabeled target domain. This process involves using the model to predict labels for unlabeled data and filtering for high-confidence samples to serve as Pseudo-Labels or proxies for the ground-truth [10]. The efficacy of this approach is further validated in recent studies, demonstrating that cyclically retraining on these pseudo-labeled samples alone is sufficient to allow a model to align with the target distribution [25]. Unlike completely source-free approaches, the source data is retained during the update cycles to make the model more stable and prevent the risk of Catastrophic Forgetting.

Test-Time Entropy Minimization

Test-Time Entropy Minimization (TENT) relies on backpropagation to update normalization layers in deep neural networks for uncertainty reduction, without ground-truth labels. It calculates entropy loss to minimize the Shannon entropy of the model's predictions on test data. As TENT adapts deep learning-based models at test time, it is also suitable for large data streams, utilizing the underlying neural network's ability to learn from complex patterns to avoid overfitting.

2.2 Review of Existing Research

Traditional Cognitive Load detection methods

In Human-Computer Interaction (HCI) and Assistive Technology, accurate cognitive load detection, especially in real-time, plays a significant role in augmenting user experience, accessibility, and performance. In the past, conventional techniques such as self-reports, EEG-based brain activity monitoring, and skin conductance have been widely used for cognitive load estimation, where EEG is regarded as the gold standard [6]. However, these techniques tend to be intrusive, lack real-time adaptation and user-friendliness, as their use in detecting Cognitive Load can hinder user movement, task performance and accessibility. EEG is accurate but intrusive, requiring heavy calibration, and serves to be less useful in real-world applications. Similarly, physiological measures such as Heart Rate Variability (HRV) and skin conductance are not as specific and reliable as they might also respond to underlying influences such as stress and physical exertion. Traditional and existing approaches are model-specific and, thus, can only detect certain SIIDs during mundane or nontrivial activities, oftentimes requiring manual tailoring of the models to fit the context. Moreover, the past constraints include the unavailability of real-time adaptation and context awareness in the systems [11].

Analyzing Eye Movement metrics for Cognitive Load estimation

To overcome these challenges, dynamic and non-intrusive eye-tracking techniques have been utilized in recent studies to provide effective indicators of cognitive load with high responsiveness and minimum disruption to the user [26].
Metrics such as fixation duration, pupil dilation, saccades, and blink rate analysis via eye-tracking techniques provide vital discernment of cognitive load assessment. This method was supported by a study that established pupil dilation as an effective indicator of mental strain in HCI [7], for it correlates with the difficulty of the task and working memory load. They have corrected the distortions in pupil-size data due to variations in gaze-angle using a neural network-based calibration model. Results from the eye-tracking experiment showed that pupil dilation was able to provide a clear indication of cognitive load changes following the calibration, thereby establishing its potential in usability evaluation. Furthermore, a groundbreaking study highlighted the potential of eye-tracking as a standalone tool for cognitive load detection, with EEG providing no significant improvement, and it comprised a dataset of synchronized eye-tracking and EEG recordings, where models based solely on eye-tracking achieved a high accuracy [26]. Additionally, a generic camera-based setup for blink detection to gauge cognitive load demonstrated comparable accuracy to a scientific-grade eye tracker, achieving 85% accuracy across different lighting conditions and an average discrepancy of 1.54 blinks each minute between two systems across task conditions [30]. Therefore, this confirms eye-tracking as a reliable, cost-effective alternative for cognitive load assessment in HCI and accessibility, particularly in video-based systems such as Augmented Reality (AR), Extended Reality (XR), wearable IoT devices, etc. Principles and importance of effective feature extraction and representation in understanding and recognizing human activities can be applied to eye-tracking data analysis, enabling an enhanced and accurate cognitive load assessment in HCI and accessibility contexts [9].

The Generalization Gap in Existing Cognitive Load Estimators

While Deep Learning methods excel in generalizing complex tasks by leveraging powerful feature extraction, they face challenges regarding large-scale data labeling dependencies, loss-function selection, and component balance [45]. Consequently, recent studies [45] highlight the urgent need for developing more robust domain adaptation methods capable of extracting transferable knowledge from limited data and effectively processing Cross-Domain heterogeneous data to ensure streamlined application in diverse fields. This necessity is further emphasized in another study [22], where the lack of generalization is identified as the primary bottleneck restricting physiological computing from real-world deployment. Key metrics, like pupil diameter, are heavily confounded by the Pupillary Light Reflex (PLR) and individual anatomical differences, often necessitating relative baselining for calibration purposes, to separate the cognitive signal from environmental noise [22].
Physiological expression of CL shifts depending on the nature of the activity; as a result, models trained on one task modality often fail to generalize to another due to the variation of specific autonomic reflexes. Studies validate that this requires a Cross-Domain Adaptation approach so that it is capable of learning the underlying, invariant features of CL that persist across disparate task environments and contexts [22]. Furthermore, real-world deployments lack the luxury of manual user-calibration phases to account for the drifting physiological baselines due to factors like fatigue, circadian rhythms, etc. This renders such systems impractical for continuous usage, thereby validating the need for dynamic adaptation frameworks that can automatically adjust to these temporal variations in real-time, as static calibration can not resolve this generalization gap. Consequently, existing CL classifiers remain person-specific or task-specific, largely failing to maintain performance when transferred to users without explicit recalibration or to new contexts. Hence, a critical gap remains for self-supervised frameworks that can dynamically adapt to diversified sensor data and shifting environmental conditions.

Evaluating Eye Movement for Human Activity Recognition

In the experimental setups for analyzing SIIDs through wearable egocentric devices such as Human I/O, a combination of egocentric video feed and microphones was captured. A consumer-level Logitech C930e compressed at 640x480 resolution, along with its integrated microphone, was used, achieving a decent set of distinguishable performed actions and situations (82 situations without overlaps) [36] while remaining replicable as a whole setup.
Another study [18] reinforces why the fusion of egocentric video & audio can achieve astonishing accuracies in activity recognition, compared to single sensor-based readings (up to 13% improvement in top-1 accuracy). Through stacking horizontal & vertical optical flow with the RGB camera data in a spatial and a temporal ConvNet, later fused with audio-based HAR via a temporal binding network (TBN), both modalities are in use.

However, recent extended reality headsets such as the Meta Quest Pro, Magic Leap 2 & Apple Vision Pro, and prototype models such as the Meta Orion & Snap Spectacles, as well as upcoming alternatives from XReal, all heavily lean on eye tracking interfaces in their design. Hence, in both wearable & extended reality devices, this can not only be a gateway to analyzing cognitive load unobtrusively as reviewed but also add another dimension to activity recognition for a more robust system. Gustafsson [15] explores this idea by attempting to classify human activities in a typical daily routine using eye tracking only. While it is promising to solely depend on eye tracking as a classifier, accuracies of up to 30-40% were reached. By itself, that would not be reliable to use; however, a multi-stack view of results combined from RGB egocentric video, optical flow, microphone input, as well as eye movement captures, not only fortifies outcomes from previous studies but also tackles the challenges of using each modality on its own. Namely, activities taking place out of frame, silent or still activities remaining indistinguishable, activities with a similar wave capture, can be better dealt with in tandem by relying on multiple modalities as inputs rather than selectively choosing one or any of their combinations.

Human I/O as the dynamic framework for SIID detection

Human I/O, a real-time system designed to enhance user accessibility by detecting a wide range of SIIDs [36], by employing a four-level detection scale to measure the availability of human input/output channels: available, slightly affected, affected, and unavailable. Information is utilized from the environment as input through senses such as vision, hearing, and touch, and communicated as an output using methods like speech, eye movements, and hand gestures in Human I/O. The authors focus on overcoming the limitations of such existing systems that fail to account for the diverse nature of SIIDs by proposing a dynamic and scalable framework where Human I/O combines egocentric devices, multimodal sensing, and large language models in predicting the user's channel availability, which noticeably improves user experience in the presence of SIIDs. The performance evaluation of Human I/O was on the basis of 300 clips selected from 60 real-world egocentric video recordings, covering 32 distinct scenarios, and the Ego4D dataset was leveraged for egocentric environmental analysis. Direct sensing techniques were also used to collect a more comprehensive set of data, such as hand detection, environmental brightness, volume level, and audio classification. The study employed a combination of computer vision, audio analysis, and large language models utilizing chain-of-thought prompting to predict channel availability with an accuracy of 82% & 0.22 mean absolute error. However, the participants wanted to draw special attention to the distinction between the necessity and desirability of these customizations as per personal preferences as well, highlighting yet another aspect of flexibility that the system stands in need of. It is proposed that combining activity and environment-based approaches, along with visual cues in tracking eye movements, would grant this system a comprehensive context-awareness of the user's situation as the mentioned system struggles with the detection of attention and cognitive load-based impairments.

2.3 Summary of Key Findings

While traditional methods like EEG are considered the gold standard for CL estimation, they remain impractical for real-world deployment due to invasiveness and hardware constraints [26]. Consequently, eye tracking has emerged as a sufficient, cost-effective, ubiquitous, and non-invasive method of CL assessment, therefore presenting an alternative to EEG [26]. Eye tracking has good accuracy under varying light conditions in most cases, but is subject to motion artifacts and calibration issues [8]. Another study [31] resolved these problems while building a low-budget eye tracker that resulted in an accuracy of 78%, performing on par with a scientific-grade eye tracker. All of these collectively signify and prove that, although not as high in accuracy as its high-end counterparts, eye tracking can serve as a logical & standalone solution for CL detection. However, a critical finding is that existing eye-tracking implementations for CL estimation are severely hindered by the generalization gap [22]. Physiological metrics like Pupil Diameter are affected by the Pupillary Light Reflex (PLR), anatomical variability, and external noise. As a result, static models trained on one task or subject fail to generalize to another, rendering generic or universal solutions ineffective while requiring a tedious calibration phase that disrupts user experience in dynamic environments like Extended Reality (XR).

To mitigate this, the findings validate that a Cross-Domain and Test-Time Adaptation approach is required for enabling dynamic adaptation. Here, it is crucial to distinguish between Domain Generalization, Domain Adaptation, and Test-Time Adaptation. While Domain Generalization constructs a static model robust enough to handle unseen distributions, the high magnitude of inter-subject variability in physiological baselines renders static solutions inefficient, as the model cannot access target information to adjust its decision boundaries [22]. Domain Adaptation addresses this by aligning distributions, but it requires access to the target data during the training phase, which is often impossible in real-time applications where users change dynamically [39]. In contrast, scenarios where a continuous stream of incoming unlabeled data from the target environment is available are better served by a Test-Time Adaptation approach. Unlike Domain Generalization, which ignores target data, and Domain Adaptation, which requires it during training, TTA actively leverages this accessible & unlabeled data stream during inference from the target space to statistically align the model's decision boundaries with the specific user's unique physiology, achieving adaptation on target data along with inference [39] [45]. Furthermore, to account for temporal instability caused by fatigue & circadian rhythms, TTA techniques can be leveraged via Iterative Pseudo-Labeling for a static artifact to transition into a dynamic system with robust performance and automatic self-calibration capabilities in real-time using high-confidence predictions from the user's incoming data [39].

These advancements directly address the identified limitations within existing SIID detection frameworks like Human I/O. The Human I/O framework lacks a predefined mechanism to differentiate between the severity or type of impairments based on the cognitive load of the user as a separate dimension alongside the existing visual, audio, and haptic cues to detect SIIDs in real-time [36]. Situations do not always provide a binary impairment; it sometimes merely delineates the availability of channels in a variable way, depending on the task itself and the amount of mental effort it requires. A gradient of interfered availability sensing would better characterize a given state of availability from the user's perspective. Hence, this finding of adapting eye-tracking as a proper & dynamic method for CL estimation would allow an efficient integration of this crucial missing CL dimension in a multi-modal sensing framework for accurate SIID detection. By integrating this self-supervised CL estimation pipeline, modern SIID detection frameworks like Human I/O can finally move beyond simple binary decisions to intelligently classify the subtle states of a user, such as distinguishing between a user who is physically occupied versus one who is mentally overloaded, thereby enabling the context-aware, nuanced adaptivity originally envisioned for such systems.

Chapter 3

Requirements, Impacts & Constraints

3.1 Final Specifications and Requirements

Cognitive load estimation based on eye tracking fundamentally depends on the availability of stable and sufficiently granular gaze signals. Prior studies consistently assume access to mid to high-frequency eye-tracking data acquired under controlled conditions, enabling the extraction of fixation dynamics, pupil responses, and fine-grained gaze behavior. In this research, eye-tracking data are drawn from COLET and ADAbase, which provide binocular gaze signals with a sampling frequency ranging from 240Hz to 250Hz. In the COLET dataset, binocular gaze signals were recorded using a head-mounted Pupil Labs eye tracker [28], whereas ADABase sourced its data using the Tobii Pro Fusion eye tracker, a stereoscopic system featuring two eye-tracking cameras for the left and right eyes, respectively [29]. Beyond hardware specifications, eye-tracking-based cognitive load estimation depends critically on calibration quality & signal stability. Poor calibration introduces artifacts and noise in gaze signals, leading to unreliable pupil and gaze features. In this work, both datasets rely on post-task subjective cognitive load assessment using NASA-TLX questionnaires. Consequently, workload labels are available only at coarse temporal granularity.

After signal acquisition, a critical system-level requirement arises from the diversity of the datasets and recording conditions. COLET and ADAbase vary in terms of structure, experimental contexts, participant behavior, and eye-tracking hardware. These differences introduce systematic domain shifts in the gaze feature distribution, even for similar cognitive states. To compensate for these shifts, a domain adaptation system is required. However, the system can not assume that a model trained on a dataset will reliably generalize to another dataset without adjustment. To address this constraint, cross-domain feature alignment is required, too. This enables a statistical harmonization between the source dataset and the target dataset prior to inference. This requirement was motivated by prior studies showing performance degradation under cross-dataset evaluation. Moreover, due to the use of coarse-grained workload labels derived from post-task assessments, fine-grained supervision at deployment time is unavailable. This is why test-time adaptation mechanisms that operate without labeled target data are required, allowing incremental model adjustments during deployment time under distribution shift [39].

3.2 Societal Impact

Cognitive load estimation has significant societal relevance in safety-critical and performance-sensitive domains where excessive mental workload can impair human attention, reaction time, and decision making. Prior research indicates that heightened cognitive load is associated with delayed response, increased likelihood of errors, and reduced situational awareness in dynamic and complex environments like human-machine interaction, driving, and aviation. Studies on driving simulation showcase that higher mental load results in significantly longer response times, reducing the driver's ability to respond to unexpected events [33]. Systems capable of estimating cognitive load from eye-tracking data can contribute positively to public safety and system usability by enabling adaptive interventions. For instance, real-time workload monitoring can support driver assistance systems that issue warnings, adjust information presentation, or temporarily reduce secondary task demands during periods of overload [12]. Apart from transportation, cognitive load estimation is also relevant in human-computer interaction and learning environments where extreme cognitive load can hinder comprehension, attention, and task efficiency. Eye-tracking-based cognitive load assessment allows interfaces to adapt content complexity and improve user experience and learning outcomes [34]. However, the societal benefits of cognitive load estimation are dependent on the generalizability and reliability of deployed models. Prior work reveals that models trained under certain experimental conditions fall short in generalizing to new tasks or sensing setups, resulting in degraded performance under domain shifts [23]. A robust cognitive load estimation system with domain adaptation and test-time adaptation has immense potential to become a decision support tool intended to enhance safety and usability.

3.3 Environmental Impact

Cognitive load estimation pipelines can impose environmental costs primarily through computation during model development and deployment. Sustainability in deep learning requires minimizing energy consumption and carbon emissions with practical criteria tied to energy efficiency and reduced carbon footprint through algorithmic and hardware choices [38]. This is especially a concern for large-scale deep learning models as they require substantial electricity and translate into significant carbon emissions.

In our work, the negative environmental impact is reduced by prioritizing lightweight adaptation and efficient reuse of trained models rather than constant end-to-end retraining under every condition. Although modern GPUs are powerful, they have considerable drawbacks in energy consumption, operational costs, and cooling demands, which motivates the use of efficient modeling strategies and hardware deployment decisions [38]. During deployment, the environmental impact not only depends on the model itself but also on where the inference runs and how frequently the adaptation occurs. Optimizing inference energy efficiency while maintaining real-time performance targets is necessary for practicality [20]. Thus, designing adaptation mechanisms that remain computationally lightweight supports both practical feasibility & lower cumulative compute over time.

3.4 Ethical Issues

Cognitive load estimation using eye-tracking data raises important ethical concerns due to the inherently sensitive and biometric nature of gaze information. Eye movements also constitute biometric data that can reveal one's cognitive state, emotional condition, attention, and even health status. Eye-tracking captures these states without explicit verbal interaction, which raises ethical questions regarding consent, transparency, and autonomy [32]. A major ethical concern is privacy and informed consent. Eye-tracking systems can infer workload, behavioral patterns, and attentional focus that users may not be consciously disclosing. Prior works reveal that gaze data can be used to identify inference and behavioral profiling, emphasizing the need for explicit consent and strict purpose limitation [42]. Since no data ever leaves the device in our pipeline, it addresses this concern. Another ethical concern arises in surveillance and misuse. Continuous gaze tracking can enable covert tracking of user attention and cognitive states, potentially leading to intrusive monitoring practices if deployed without clear safeguards [19]. Such risks are particularly salient in institutional contexts such as workplaces, educational environments, or public spaces. Ethical challenges are further heightened in vulnerable or sensitive settings, including healthcare and child-computer interaction. Thus, implementations require careful ethical review to avoid unintended psychological or behavioral consequences for participants [21].

3.5 Standards

This research aligns conceptually with widely recognized standards that govern human-centered system design and the responsible handling of personal data. In terms of data protection and privacy regulations, eye-tracking data falls under personal or biometric data categories. Analyses of eye-tracking technologies under the European Union General Data Protection Regulation (GDPR) emphasize that gaze data can become biometric data when used for identification, profiling, or inference of sensitive attributes, thereby invoking stricter legal protections. Compliance with GDPR and consent when collecting and processing gaze data is required [35].
As per the FAIR & CARE principles, the selected datasets, COLET & ADABase, are anonymized, and participant details cannot be traced back to individual identities [28][29].

3.6 Project Management Plan

The project lifecycle is divided into 5 crucial phases:

1. Establishing the Rationale & Background Literature Review (2-4 weeks)
2. Data collection, Preprocessing, Feature Engineering & Alignment (4-5 weeks)
3. Baseline model testing & External Validation for Domain Shift Analysis (2-3 weeks)
4. Static Correlation Alignment implementation & Dynamic Iterative Pseudo-Labeling implementation for TTA (5-6 weeks)
5. Extraction of all performance metrics & plots of relevant graphs, interpretation of the results through XAI, and identifying limitations of the implementation for future work recommendations (4-5 weeks)

Each phase is treated as one sprint, with each sprint lasting a maximum of 6 weeks, accounting for varying schedules during ongoing academic semesters.

A high-performance PC from our Research Lab was allocated for all phases involving testing, free of cost. For all data preprocessing, feature engineering & alignment, model training, and evaluation, three tiers of PC configurations were tested for validating the findings. The high-end system with an i7-14700K, RTX 4080 Super, and 64GB of DDR5 RAM was used to ensure stable initial results. These results were then reproduced on a mid-tier laptop with a Ryzen 7 8840HS, RTX 4060, and 16GB of DDR5 RAM, to represent a wider range of end-devices. Since our training & testing were done on CPU-only, to represent a CUDA-disabled, iGPU-only low-end configuration, a Ryzen 5 2400G-based desktop with 8GB of DDR4 RAM was selected for a final confirmation of the findings.

3.7 Risk Management

The primary technical risk for CL estimation is domain shift. Differences in eye-tracking hardware, participant behavior, and experimental protocols cause performance degradation when evaluated across datasets [23]. This risk is addressed using feature-level domain adaptation to reduce distributional mismatch without requiring labeled target data [14]. Another risk is noise and calibration instability. Eye-tracking measurements are sensitive to calibration quality, blink-related signal loss, lighting conditions, and head movement [12]. This is managed through careful preprocessing, normalization, and exclusion of unreliable samples. Adaptation instability is another risk that gets introduced when working with test-time adaptation. To mitigate this, selective and constrained adaptation strategies are used to prevent unstable behavior during inference [39].

3.8 Economic Analysis

Deploying cognitive load monitoring systems has direct and indirect costs that impact the feasibility in real-world settings. Practical adaptation depends on whether the devices are affordable, durable, and acceptable to users. It also depends on whether organizations perceive a positive cost-to-benefit ratio. Occupational safety and health (OSH) professionals suggest that organizations have shown a willingness to pay for wearable monitoring devices. Respondents estimated their organization would spend about $72 per person for a wearable device, indicating that per-user hardware budget may be quite modest in real deployments [16]. Costs aren't limited to purchasing devices. Additionally, recurring operational costs depend on sensor durability, privacy & confidentiality concerns, and employee compliance. These issues have a major economic impact as they influence rollout success and long-term maintenance efforts [16]. A common economic bottleneck that many solutions face is that they become non-scalable systems with the lack of interoperability, raising integration costs. These can significantly increase the cost of ownership, even if the cost of sensors is cheap [17]. For a cost-conscious design, our implementation focuses on the efficacy of lightweight yet reliable models for CL estimation.

Chapter 4

Proposed Methodology

4.1 Methodology Overview

We propose a 5-stage CL estimation pipeline as shown in Figure 4.1, which utilizes domain adaptation & test-time adaptation methods to create a dynamic, domain-adapted model that is able to retrain itself without requiring ground-truth labels.
Firstly, raw eye-tracking features related to pupil dilation, gaze patterns, and blink rates are collected from two separate datasets - COLET & ADABase. Subsequently, gaze, pupil-related & blink-related features are extracted from the raw data. To address baseline shifts in the features due to physiological changes and unit discrepancies between the two datasets, independent robust scaling is employed on the extracted features.
After the feature extraction phase, the Data Alignment & Integration step ensures compatibility among the datasets. This is done by normalizing the column names, finding the intersection of the common features, and merging the datasets into a single form, where the scores of the NASA-TLX serve as the ground-truth labels for CL. The top 30 discriminative invariant features are then selected using a Random Forest classifier to reduce the dimensionality of the data.
The heart of the pipeline is the Domain Adaptation stage. On the first pass, a HistGradientBoostingClassifier is trained on COLET & validated on ADABase, and on the second pass, another HistGradientBoostingClassifier is trained on ADABase & validated on COLET to serve as the baseline metrics. However, due to the performance degradation that may occur from the domain shift, Correlation Alignment (CORAL) is used to align the covariance of the features in the source data with the target data, ADABase.
Lastly, to ensure the model is robust in real-world deployment scenarios, we use Test-Time Adaptation (TTA) in the form of iterative pseudo-labeling. In this phase of the process, the model is used to make initial predictions on the unlabeled live target stream. The high-confidence predictions ($> 85\%$) are then filtered and used as pseudo-labels to iteratively retrain the model via self-training. The output of the pipeline is a classification of the user's cognitive state into one of the four classes: Very Low, Low, High, or Very High Cognitive Load.

A diagram illustrating the proposed methodology:

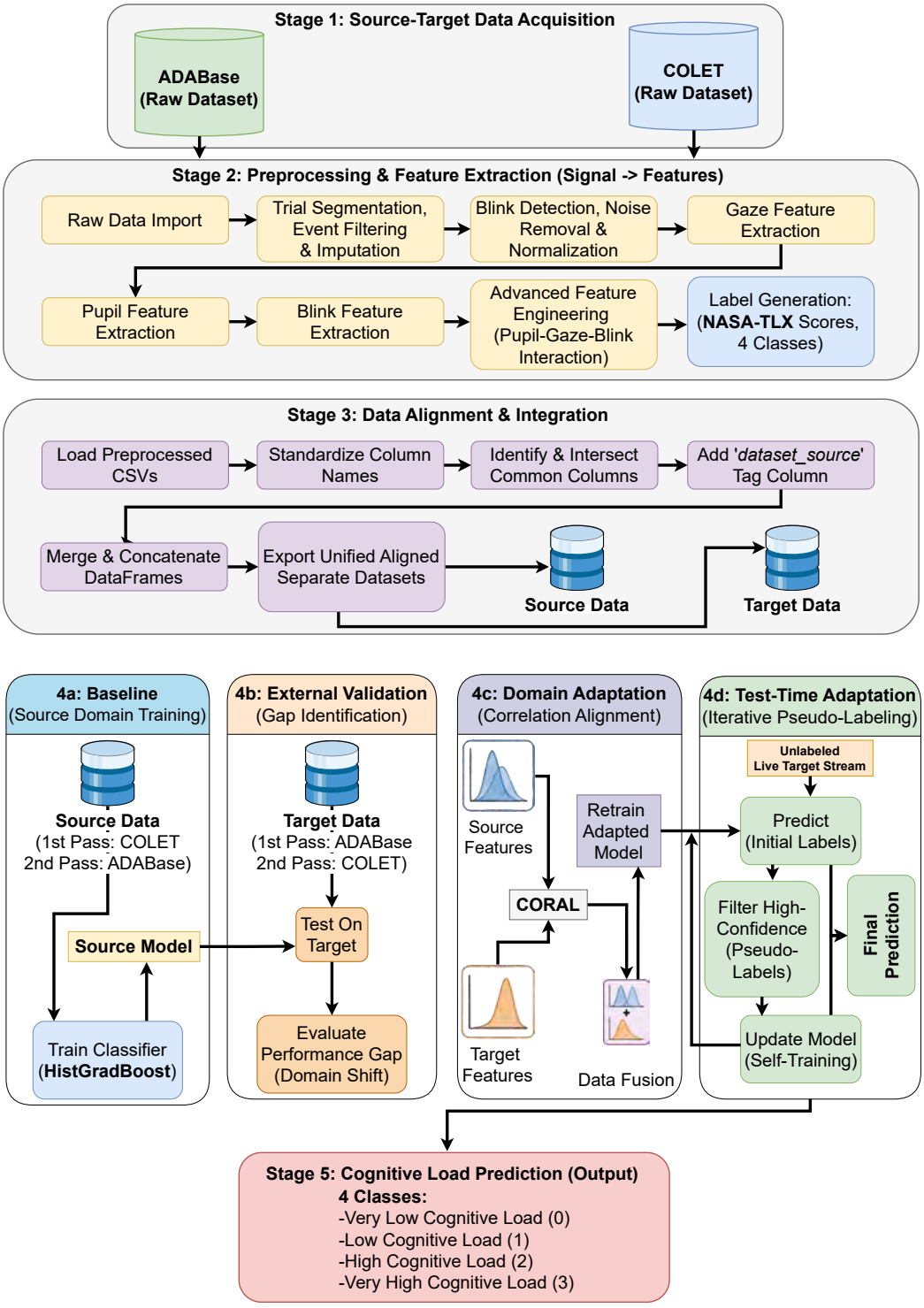

Figure 4.1: Complete pipeline of the proposed CL estimation model

4.2 Model Specification

4.2.1 Baseline Comparison

To create a solid foundation for the baseline scores, we have selected a HistGradientBoostingClassifier, which is an ensemble model based on decision trees. This estimator was chosen for two important reasons. Firstly, it natively supports missing values without the need for aggressive imputation; this is especially common in eye-tracking data due to blinks & artifacts that can introduce bias. The second important reason is that it bins continuous feature values into integer-valued histograms, drastically accelerating training on tabular data while maintaining high interpretability.

A maximum of 200 boosting iterations was used to compensate for a slower learning rate of 0.05, allowing the model to converge gradually without memorizing sensor noise or stopping prematurely. A maximum tree depth of 5 was set to cover the non-linear interaction between the pupil and gaze features, as a higher depth would make the decision tree too complex for lightweight implementations. In addition, L2 regularization was used to prevent the model from overfitting the source domain by setting `l2_regularization = 0.1`. A random seed of 42 was also set to ensure the reproducibility of the domain adaptation results.

A ResNet-MLP architecture was used to compare the HistGradientBoostingClassifier to DL-based pipelines. The Residual Multi-Layer Perceptron with two residual blocks and a linear classifier was trained with Batch Normalization and the Dropout, p set to 0.3.

4.2.2 CORAL

When testing the model on an external dataset, it can encounter diverse target domains. In these settings, domain adaptation techniques can minimize the bias induced by the training dataset [40]. In order to reduce covariance mismatch during adaptation, we leverage Correlational Alignment (CORAL) to account for subjectivity. By aligning covariance matrices in all domains, CORAL can minimize the shift induced by the fact that NASA-TLX scores may differ from one subject to another subject. Since CORAL needs to be trained on a model, the baseline HistGradientBoostingClassifier model will be reused with the fine-tuned hyperparameters for comparison.

4.2.3 Iterative Pseudo-Labeling

For Test-Time Adaptation, the static CORAL-adapted HistGradientBoostingClassifier was reused with the same parameters as before. The confidence threshold was set at 0.85 as a balanced sensitivity, without making the model too flexible or too rigid. The iteration count was set to 10 in order to stop the model early, so that the model does not become overconfident. However, both of these parameters have been evaluated over a wider range in section 5.3 to validate the chosen optimal values.

4.3 Data Collection

Multimodal datasets containing both eye-tracker information and ground truth labels for cognitive load metrics are sparse and domain-specific. COLET, or Cognitive workLoad estimation based on eye-tracking, is an open dataset that bridges that gap by monitoring 47 independent participants via Pupil Core eye-trackers, as they were tasked with solving visual puzzles of varying difficulty. The raw data were compiled from multiple sources per trial as one task was performed by a subject. The collected data streams included pupil data for continuous measurements of the diameter of both pupils, gaze data for X and Y coordinates of eye gaze position over time, blink data to represent the times of occurrence and durations for the blink, subject information, annotations, and metadata. Participants were surveyed at the end of each activity by having them fill out a NASA-TLX questionnaire, which assesses the workload subjectively. Since EEG-based physiological measurements are not applicable to end devices with limited sensing capabilities, subjective measurements have proven to be acceptable for categorizing cognitive load [28].
In accordance with our proposed Source-Target Domain Adaptation pipeline, we utilized ADABase as the designated Target Domain. Unlike a traditional external validation set, this dataset serves a dual purpose: it acts as the unlabeled stream for our Test-Time Adaptation (TTA) experiments and as the target distribution for Correlation Alignment (CORAL). To enable these cross-domain tasks, the raw ADABase data required a specialized ingestion pipeline (Stage 2) followed by a rigorous alignment phase (Stage 3). The raw ADABase dataset was acquired as 30 distinct HDF5 containers, which were unpacked into a structured directory of subject-specific folders (for example: adabase-public-0001) [29]. Each folder contained synchronized CSVs for signals, metadata, and subjective scores. Unlike the discrete task files in COLET, ADABase presented continuous signal streams. We implemented a Trial Segmentation process that iterated through the signal files, grouping continuous data points based on unique study and level identifiers in the metadata to reconstruct discrete experimental trials.
Usage permissions have been obtained via a signed End-User License Agreement for the legal utilization of each dataset required. As per the FAIR & CARE principles, the datasets are reproducible and interoperable. They are also anonymized, and subject details cannot be traced back to personal identification factors.

4.3.1 Exploratory Data Analysis

First, we performed exploratory data analysis on the COLET dataset. To break down the complexity, the raw data was transformed from its original .mat file type to standalone CSV files, as shown in Figure 4.2. In total, the dataset includes 47 subjects organized as follows: each subject includes an id_info.csv file containing demographic data, and each of the four tasks (tasks 1 through 4) includes time series gaze_data.csv, pupil_data.csv, blink_data.csv, and annotation (NASA-TLX scores) CSV files. The time series files were processed such that a summary feature set for each trial could be extracted prior to modeling. Through this process, a feature table was organized for analysis with 188 unique observations.

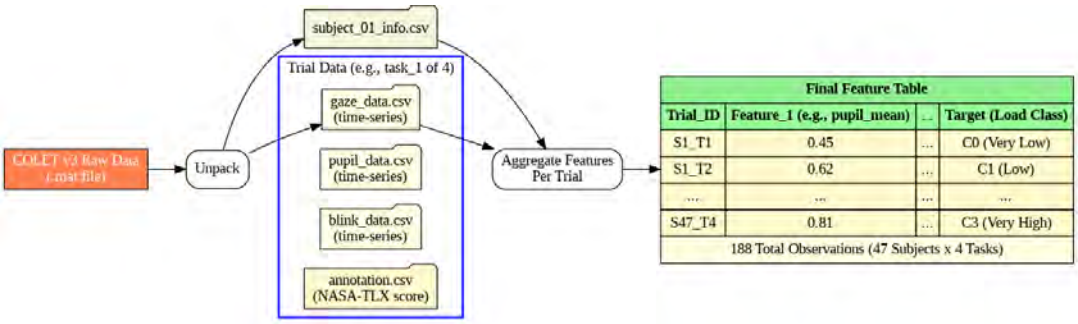

Figure 4.2: COLET Feature Table Summary

Looking at a portion of the tabular data in Table 4.1, we can have an estimation of the data types that can be encountered from the base feature set distribution:

subject_id	task_id	Gender	...	blink_mean_duration	NASA_TLX_mean
10	1	0.0	...	0.252	16.7
10	2	0.0	...	NaN	9.2
10	3	0.0	...	0.282	45.0
10	4	0.0	...	0.225	44.2
11	1	1.0	...	0.177	23.3

Table 4.1: Excerpt of the features with NASA-TLX mean workload scores as the target (middle columns omitted for brevity).

All categorical data was encoded accordingly, thus no further transformations are required on the domain of existing numerical data types.
As null values were present, particularly within the features associated with blinking, they require addressing when proceeding with the data cleaning segment.

Observing the correlation heatmap of the dataset in Figure 4.3:

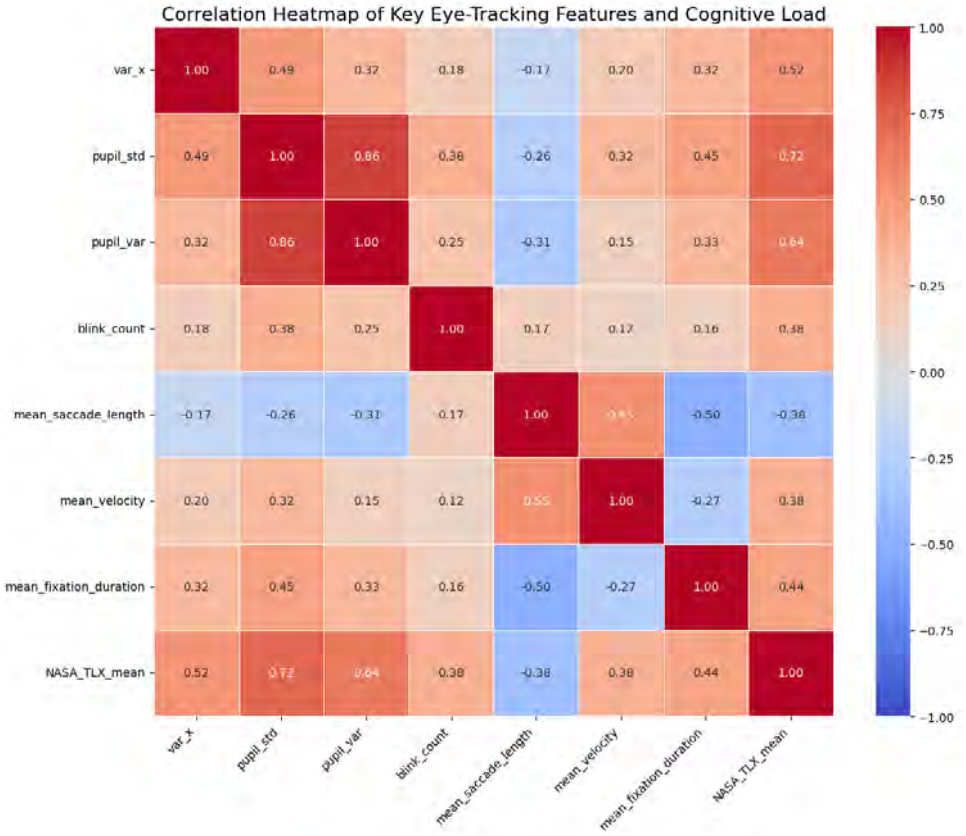

Figure 4.3: COLET Correlation Heatmap

The features - standard deviation of pupil diameter & the variance of pupil diameter over the 1s time window of each sample seem to have a strong positive correlation with each other, as well as with the mean NASA-TLX scores. On the contrary, the mean fixation duration per participant has a moderate negative correlation with the mean of the recorded amplitude of saccades, as observed in Figure 4.3.

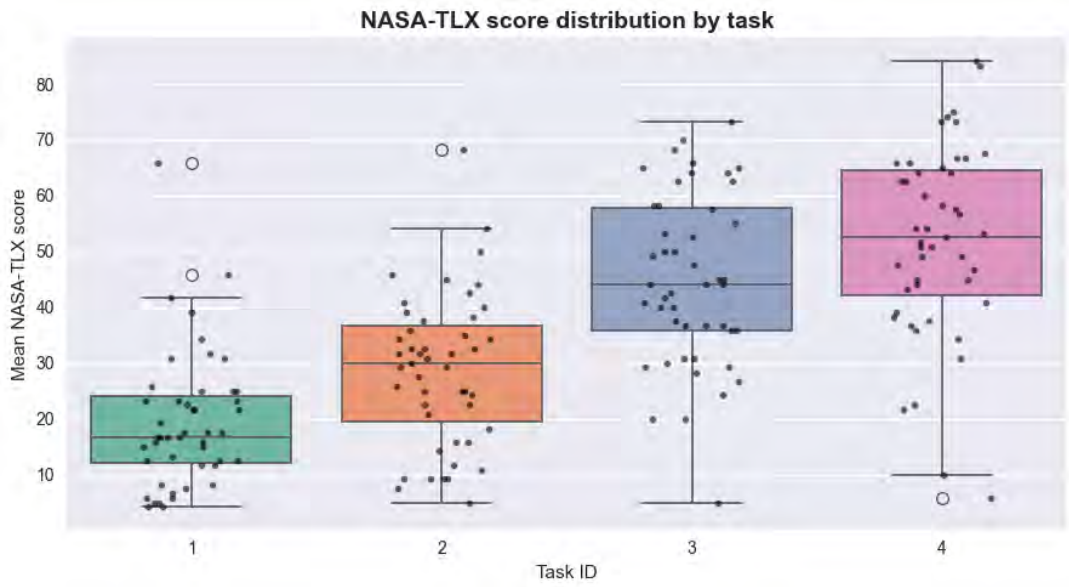

Figure 4.4: COLET Task-wise NASA-TLX Mean

Mean NASA-TLX score distributions for each task are as seen above in Figure 4.4. Task 1 contains overall lower workload scores, with most participants reporting scores between about 10 and 25 and mild outliers above 60. Task 2 shows a moderate amount of workload, where scores are reasonably distributed between about 10 and 50. Task 3 and 4 show moderate to high perceptions of workload, with respective median scores nearing 45 and 52, with both showing greater variability than the other tasks. In particular, outliers were evident in every task, which meant some subjects experienced relatively low or high workloads compared to the majority of subjects, particularly in task 1 and 4.

Next, prior to the utilization of ADABase as the target domain, an exploratory analysis for this dataset was also carried out in order to affirm the quality level of the mapped features as well as the distribution level of the subjective ground truth. This analysis played a crucial role in confirming that the target domain contains detection patterns similar to the source domain (COLET).
To validate that the ADABase tasks induced distinguishable levels of cognitive load, we analyzed the distribution of NASA-TLX scores across the different experimental tasks. Figure 4.5 illustrates the spread of subjective workload ratings for each valid task identifier.

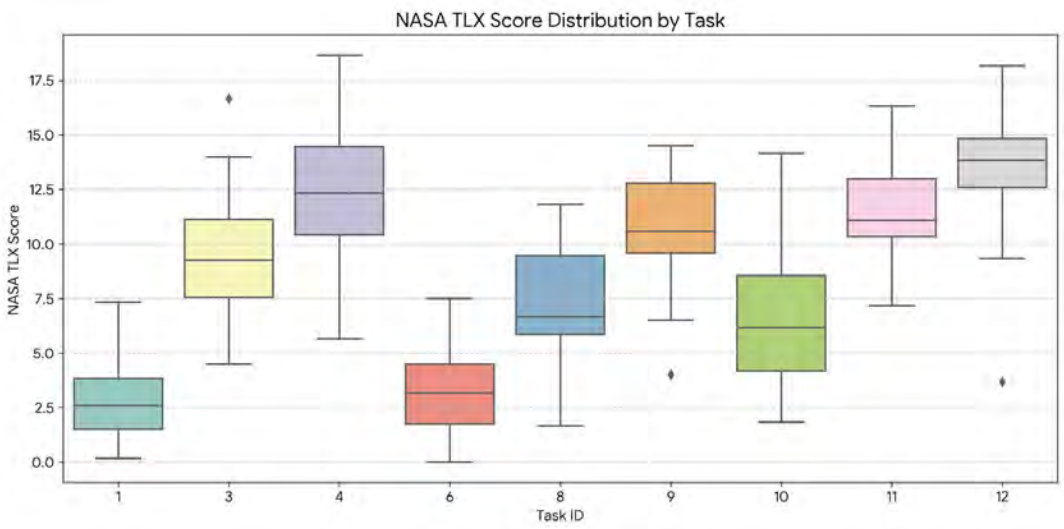

Figure 4.5: ADABase Task-wise NASA-TLX Mean

As can be seen in Figure 4.5, a lot of variability between the tasks is present, which again confirms the observation that the data set indeed represents a wide spectrum of cognitive states from low to very high levels of cognitive effort. What is important to highlight from the above results is that the information from Tasks 2, 5, and 7 was not considered for this particular assessment pipeline, leading up to the actual training pipeline, because they did not have the relevant ground truth provided through the NASA TLX labels in the raw data set provided.
In order to understand the correlation between the derived physiological features and the cognitively labeled data, we developed a correlation heatmap. Figure 4.6 shows the important features related to oculometrics and highlights the correlation.

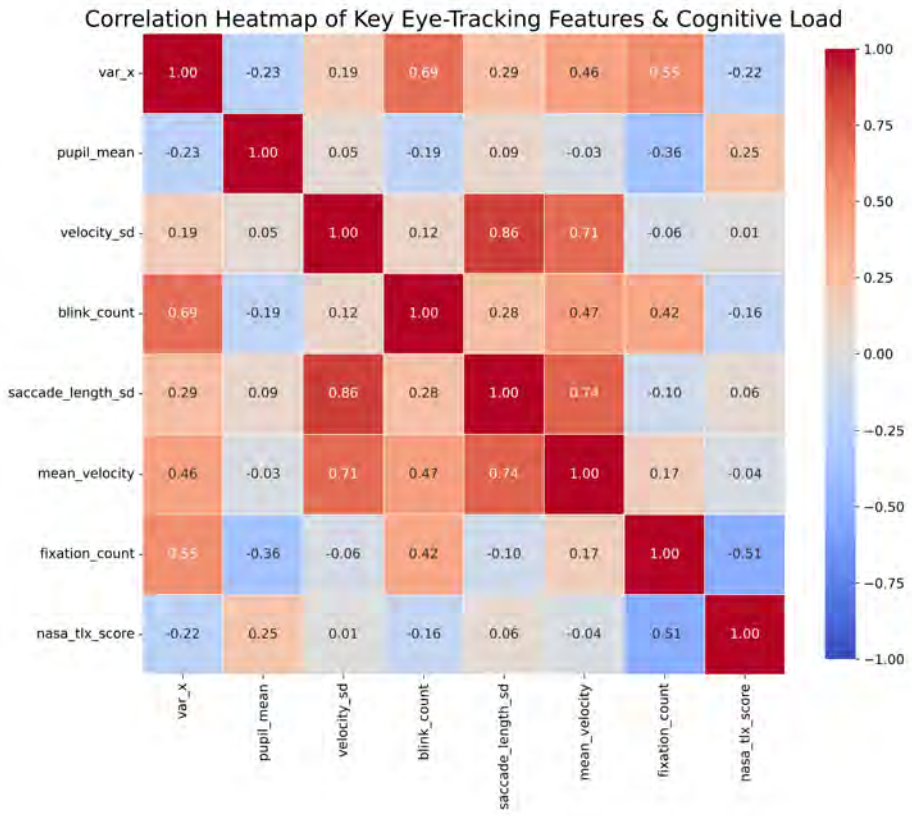

Figure 4.6: ADABase Correlation Heatmap

Similarly, akin to the findings from the source domain, there were physiological changes to higher workload. Features related to the pupils' dimensions, specifically the average pupil diameter and its robust variations, demonstrated a positive correlation in line with the task-evoked pupillary response. According to the Task-Evoked Pupillary Response (TEPR) theory, pupil dilation acts as a reflex to cognitive load [2]. Additionally, a considerable variance was observed in blink frequency and duration compared to varying load intensity, thus justifying their inclusion in our feature set. The heatmap above also shows blocks of highly correlated features such as the average speed of eye movements (`mean_velocity`) and the variability in the distance of eye jumps (`saccade_length_sd`). This implies a consistent physiological signal in the feature set from varying sessions of the target domain.

4.3.2 Data Cleaning

The raw eye-tracking time series data required initial preprocessing as it was inherently noisy. It was achieved through converting the raw data into a structured format by calculating 33 base features containing robust statistical summaries (e.g., mean, standard deviation & velocity for pupil, gaze, and blink signals), which resulted in effectively smoothing the data and preparing it for subsequent feature engineering, as meaningful stable signals have been extracted from noisy raw inputs. Afterwards, the missing values were analyzed on the initial COLET feature set. The analysis indicated that data sparsity is still a limitation, affecting only blink-related metrics.

- blink_total_time: Null count 40
- blink_duration_sd: Null count 40
- blink_rate: Null count 40
- blink_mean_duration: Null count 40
- blink_count: Null count 40
- blink_frquency: Null count 40

The reason for the missing features is that some subjects did not blink during the considered window duration. To exemplify, Subject 10 did not blink at any given moment while completing task 2 within about 29 seconds. Hence, this resulted in an equal number of missing entries for all blink-related features.
We used a multi-stage approach to handle missing values like these to ensure data integrity and leakage prevention. In the initial feature extraction phase, each signal extractor incorporated a built-in method for flagging missing sensor data (e.g., `blink_feat_missing` = 1.0) each time the corresponding raw channel was missing. This way, we retained information about missing sensors while keeping all features numeric. The other metrics derived from gaze and pupil data were complete.
In the primary imputation phase, after feature_cols had been selected, the dataset was filtered using column-wise median imputation (X = X.fillna(X.median())) and subsequently used zero imputation (X = X.fillna(0)) for any feature or voiding columns. The train/test split was performed before augmentation, and imputation was handled separately to avoid data leakage.

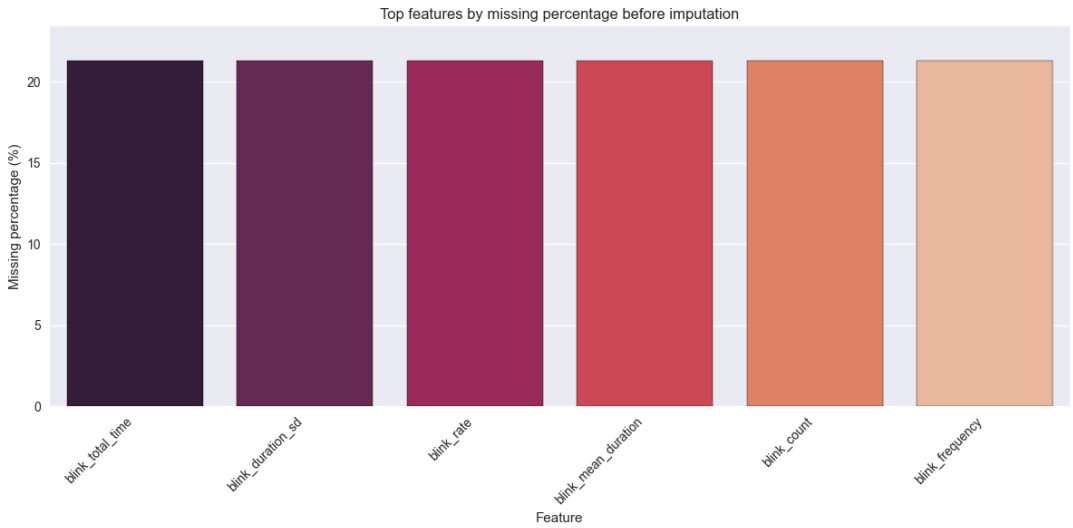

Figure 4.7: Missing Features' Plot of COLET

For higher-level meta-features such as subject or task aggregates, any remaining missing entries were replaced with global subject or task means calculated from the training set. On the complete data for engineered features, all remaining NaN or infinite values were replaced with zero, and then the VarianceThreshold filter was applied to remove constant or near-constant features to keep only finite and informative variables. In all cases, outliers and inconsistent values were addressed

using a combination of robust statistical methods, explicit missing-value indicators, and careful imputation steps. For extreme values, a combination of IQR clipping, RobustScaler, as well as logarithmic or exponential transformations was used to stabilize the extreme values. All of these steps raised robustness during modeling and maintained data integrity in the imputation process. As a result, outliers and inconsistencies were addressed effectively.

Contrary to post-hoc statistical cleaning that was performed in COLET, data cleaning for ADABase was actually incorporated into the signal ingestion pipeline due to its nature. An inferential logic was built into the system to identify the valid signal code (0 or 1) for a specific subject. The raw pupils and gaze sample values marked as invalid based on the `LEFT_PUPIL_VALIDITY` and `RIGHT_PUPIL_VALIDITY` columns were set to zero before the feature extraction phase to avoid sensor noise.
To overcome the issue of sensor dropouts in a single channel of the data, the nan-mean of left and right channel data was computed in a similar way like Figure 4.8. This actually smoothed the data and filled the spots where one of the eye's data was not available.

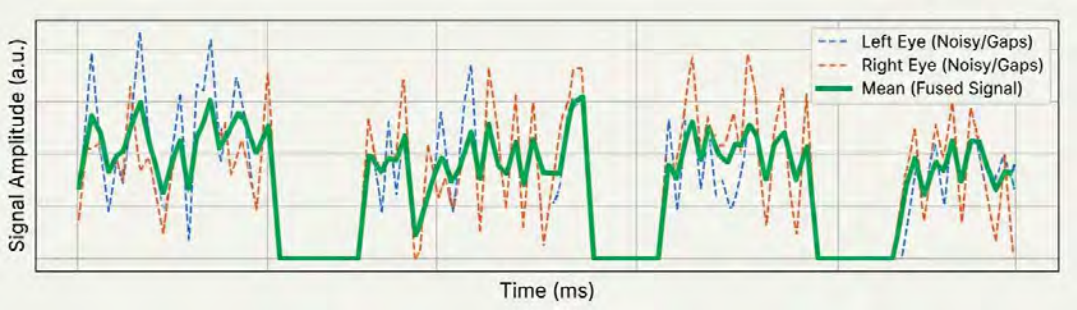

Figure 4.8: ADABase Pupil Data Signal Fusion

To maintain strict quality control over our ground truth labels, we filtered out all trials in which essential information, such as Study/Level identifiers, was missing. Additionally, all those sets of tasks that did not correspond to all the NASA TLX subjective load values were removed to ensure all data had a cognitive load label assigned to it.

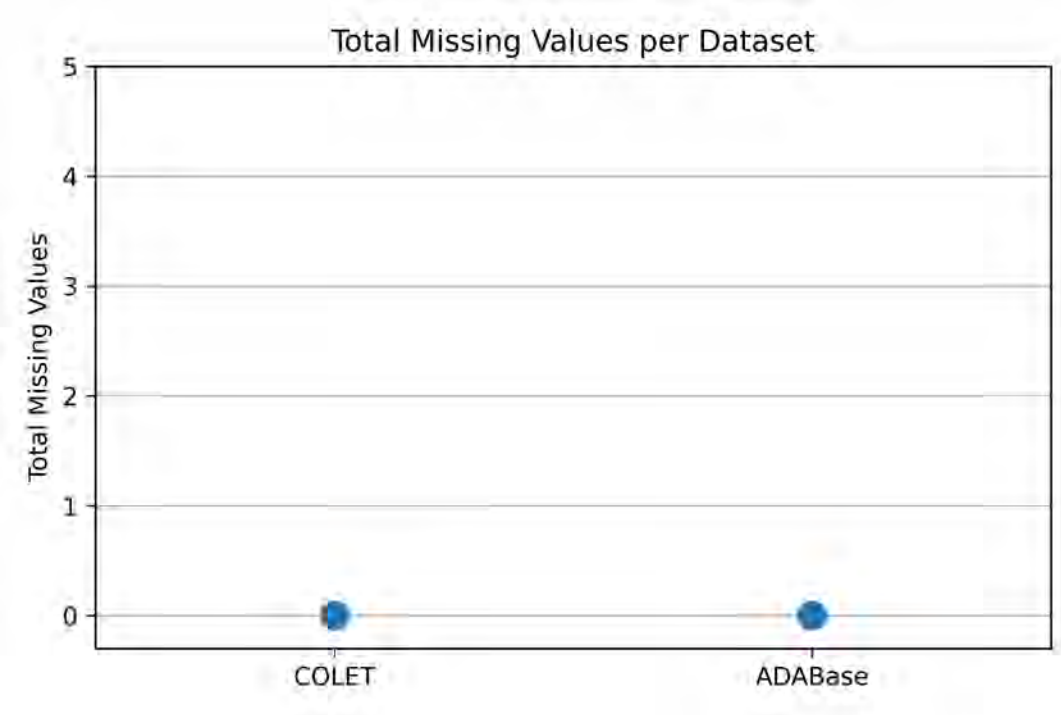

Figure 4.9: Missing Features' Dot Plot after Data Cleaning

To demonstrate the effectiveness of the data cleaning stage, a dot plot has been visualized, showing an absence of missing value trends through the samples, as depicted in Figure 4.9.

4.3.3 Data Transformation

In order to model the physiological complexity of cognitive load correctly, the initial set of 33 basic oculometric feature measures was insufficient for effective classification. Therefore, we increased the feature set to 94 sophisticated attributes through the successful employment of a stringent feature engineering process backed by recognized psychophysiological scientific literature. This was important for correctly dealing with the non-linearity between eye measures and mental operations.
From the raw signal distribution analysis, high levels of right-skewed tendencies were evident in the key parameters of interest, e.g., the pupil diameter and fixation time. Normalization of the distribution and the detection of underlying non-linear patterns in the data were attained using logarithmic and polynomial transformations of the base features, specifically, squares and cubes, thereby being similar to the standard approach that various sources support for the handling of physiological reactions that tend towards log-normal distributions [4].
Beyond statistical transformations, we have engineered specific composite indices that were designed to function as direct proxies for cognitive states. It is computed by dividing the pupil diameter by the velocity of the saccades ($\frac{Pupil}{Velocity}$) [5].
The feature is supported by the capacity model of attention. According to this, when the demands of a task increase, sympathetic arousal (reflected by pupil dilation) and tunnel vision (low velocity of saccades) both tend to increase [1] [2]. Defined as the fixation count over the total saccade distance. It evaluates the user's

visual concentration level. This measure uses the terms focused processing in relation to high values, while presenting distracted scanning in relation to lower values. We also generated interaction features that model the rich dependencies between different oculomotor systems. These were obtained by multiplying the pupil diameter by the coordinates of gaze, thus capturing the gaze-dependent pupil response and removing the cosine error, where the apparent size of the pupil changes with the angle of observation with respect to the screen center [13]. We also included highly discriminative ratio features, such as blinks per fixation, that capture the trade-offs between intake of visual information through fixations and cognitive breaks through blinks, hence offering to the model a more context-enriched representation of the subject's mental state [5].

It has also been discussed that in order to detect such entities as SIID (Situationally Induced Impairments and Disabilities) appropriately, it is not enough to consider a binary or ternary classification to create an adaptive system. Following the taxonomy of user availability as defined in the Human I/O framework, which has four classes of user availability: Available, Slightly Affected, Affected, and Unavailable, our model of cognitive load has also been developed to predict it as a 4-class problem (C0 - C3) [36]. This allows differentiating between users being busy, which translates to High Load, and users being cognitively unavailable, which translates to Very High Load.

After feature extraction, there seemed to be a significant imbalance of the target labels. More specifically, a lack of entries for very high cognitive load (C3) can be observed in Figure 4.10:

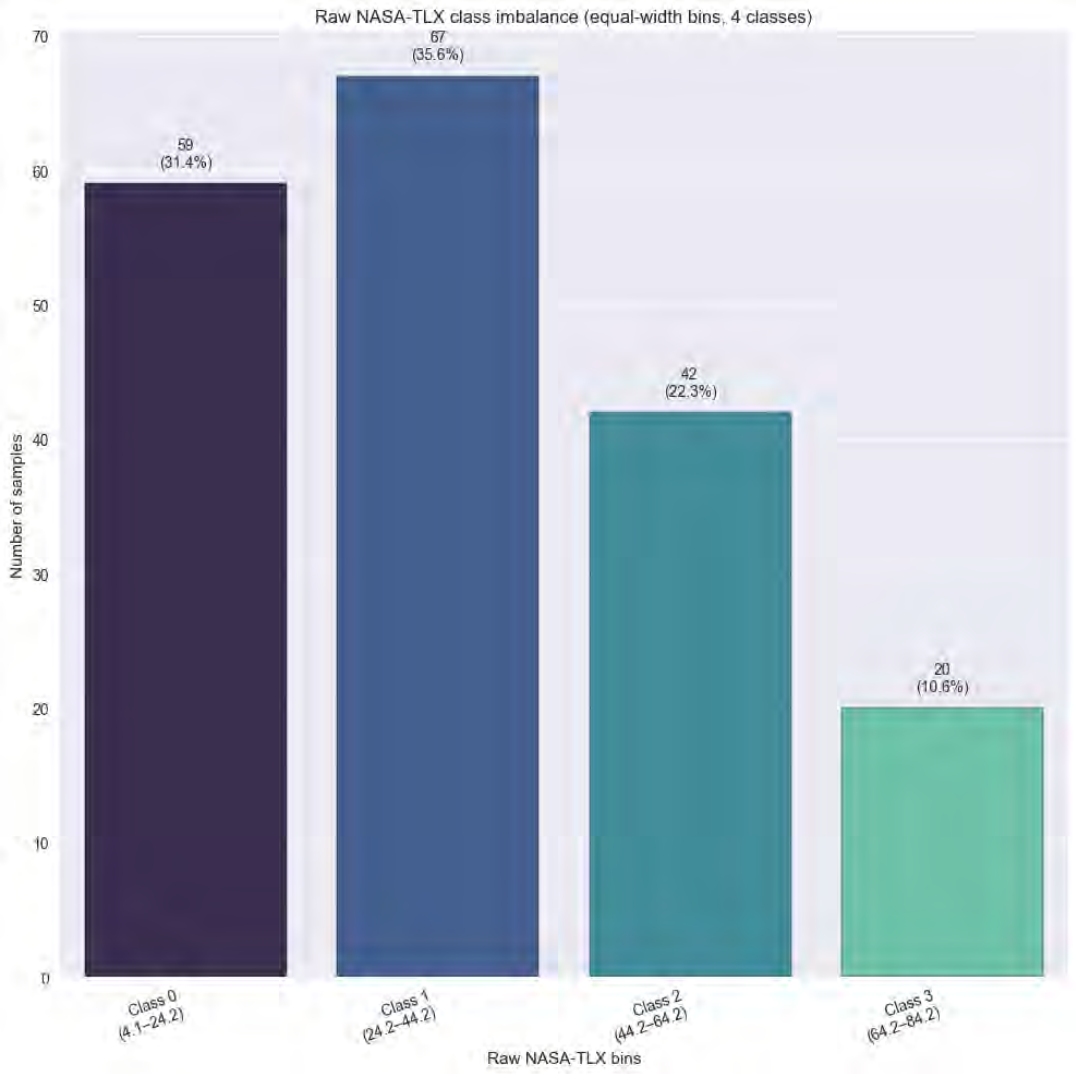

Figure 4.10: Imbalanced Target Label Distribution of COLET (equal-width bins)

Therefore, to balance the ground truth labels for COLET, normalizing the target (NASA-TLX mean scores) using a Min-Max scaler followed by determining the class boundaries using rank-based quantile distribution was performed. This ranked all the trial scores in ascending order and split them into four equal quartiles. The class boundaries derived were: Very Low CL (C0) = [0.00, 0.219], Low CL (C1) = (0.219, 0.385], High CL (C2) = (0.385, 0.573], and Very High CL (C3) = (0.573, 1.000]. Unlike traditional thresholding, which often leaves extreme classes underrepresented, this partitioning ensured that precisely 47 out of the 188 recorded scores (25%) were assigned to each class, including the Very High CL, C3 category. This efficiently neutralized the intrinsic imbalance of the raw subjective scores before the model training phase.

It is noteworthy to mention here that the data split was directly performed to ensure the training dataset consists of 80% of the total data observations, while the remaining 20% constitutes the testing set directly after the feature extraction process. This is to avoid the Lucky Split effect normally obtained while working with relatively limited physiological datasets, such as ours, where the total observations are about 200. During the train-validation split itself, Stratified Splitting (stratify=y) was employed. Thus, the proportionate spread of the four cognitive load

classes in the test dataset remained as it could have been in the original dataset in the sense that even the Very High CL (C3) class would not be compromised by the split itself. Moreover, the test dataset remained distinct from the rest of the data splitting operations, i.e., it remained untarnished during the train-validation split. Given this intrinsic class imbalance within the training set, a manual Random Oversampling strategy was preferred over any synthetic generation methods. A specialized function for balancing data was put into place in order to find out the frequency of the majority class and make sure to artificially inflate the minority classes by sampling existing instances with replacements. This was done until every count from each class matches the majority class exactly, as shown in Figure 4.11. Through setting perfect class parity before training the HistGradientBoostingClassifier, we neutralized intrinsic model biases toward the majority class. As such, the gradient boosting trees now learned the distinguishing physiological patterns for High Load versus Low Load, related to features like pupil dilation and saccade velocity, rather than just minimizing error based on prior probabilities.

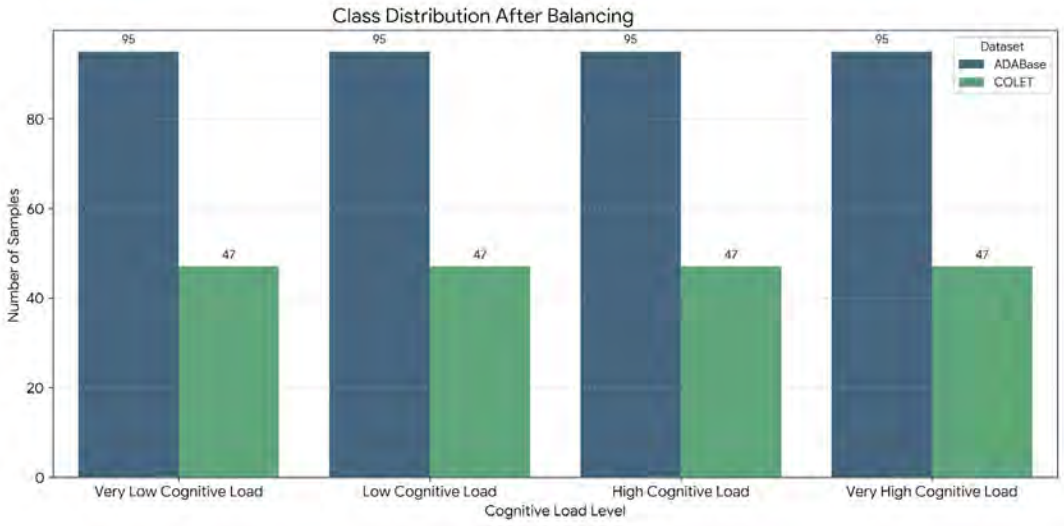

Figure 4.11: Class Distribution After Balancing

4.3.4 Data Integration

After independently preprocessing the source domain (COLET) and the target domain (ADABase) individually, a stringent data alignment phase has been executed. This phase has been carried out to ensure compatibility between the involved data schematics before the actual consumption of the machine learning models. As depicted in Stage 3 of the methodology, the data alignment phase has been conducted by performing three crucial steps:
Since the data was obtained from distinct sources, the attributes also had varied formats. To standardize the target domain in relation to the source domain, the attribute nasa_tlx_score from the ADABase standard was given the same mapping as the NASA_TLX_mean attribute from the COLET standard. Similarly, the flags for valid signal and the score were standardized to attain the same ground truth attributes, NASA_TLX_clean and NASA_TLX_normalized, for the two dataframes. To ensure there is no feature space mismatch during our Domain Adaptation phase, we mathematically calculated both datasets' feature intersection sets. Although

both pipelines provided us with rich sets of features, we only ended up using features present in both datasets. This process eliminated dataset-specific artifacts and led us to our concluding set of 94 physiological features, as both datasets contained information on pupil dynamics, blinks, and saccade kinematics. To enable the later Domain Adaptation stages (CORAL and TTA), a metadata tag `dataset_source` was appended to all trials.
This alignment operation produced two data sets that are in synchronization. This method separates both datasets into two separate csv files but keeps them compatible with the entire pipeline so that the Source-Only Training (Step 4a) and the Target-Only Validation (Step 4b) can be performed without any data leakage.

4.3.5 Summary of Preprocessed Data

After the intensive preprocessing steps, the final architecture of the data set was consolidated to two synchronized data streams for processing by the domain adaptation pipeline. It was identified that both COLET and ADABase datasets were reduced to a final intersection of 30 top behavioral features using a RandomForest classifier, as shown in Figure 4.12. The features cover major physiological areas such as advanced engineered features related to pupil-gaze-blink correlations, pupil mean diameter, variance, velocity, quantile metrics, gaze saccade length, velocity, spatial distribution (X, Y coordinates), blink Rate, blink duration, and blink frequency.

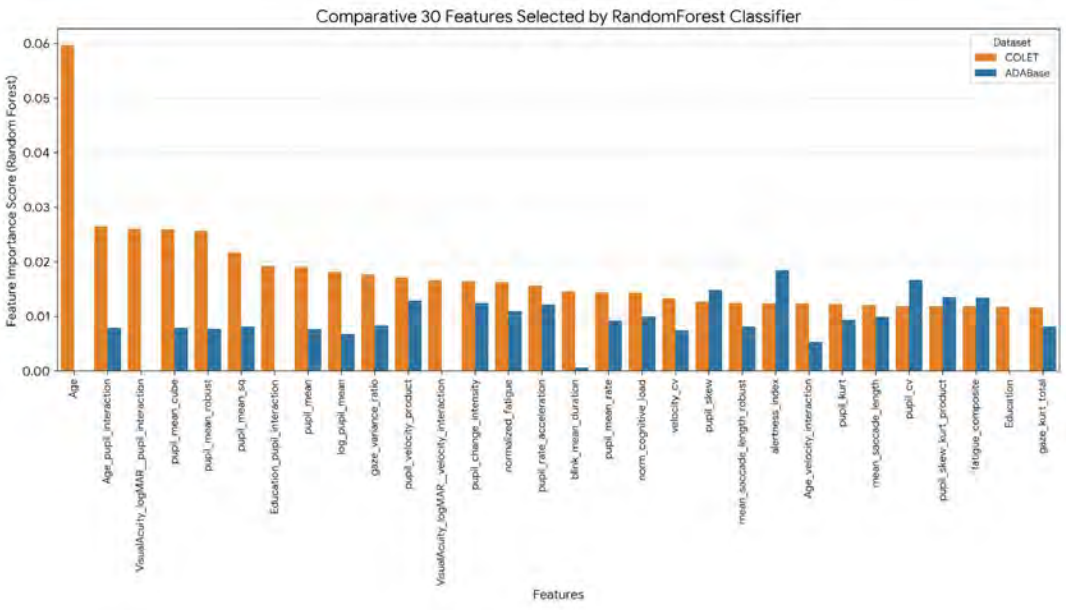

Figure 4.12: Comparative 30 Features Selected by RandomForest Classifier

pupil_mean_cube	alertness_index	...	gaze_variance_ratio	nasa_tlx_mean
12.21	0.45	...	2.10	16.7
10.55	0.52	...	1.85	9.2
15.30	0.33	...	3.44	45.0
14.12	0.38	...	2.95	44.2
9.88	0.61	...	1.55	23.3

Table 4.2: Excerpt of COLET trials showing key features

pupil_mean_cube	alertness_index	...	gaze_variance_ratio	nasa_tlx_mean
19.77	0.08	...	1.07	7.00
20.28	0.14	...	1.06	11.33
20.85	0.15	...	0.99	12.00
18.54	0.18	...	0.37	1.50
20.98	0.10	...	0.59	4.83

Table 4.3: Excerpt of ADABase trials showing key features

As indicated in the comparison of the structures in Table 4.2 and Table 4.3, the feature space of both datasets has been successfully aligned and ready to be fused in spite of their disparate original forms in the raw data.

4.4 Implementation of Selected Design

The strategy for implementing these models was to prioritize robustness over complexity while making a shift from deep-learning models to a HistGradientBoosting-Classifier to alleviate the prevalent issue of overfitting faced on limited physiological data. The most crucial part of this design was to ensure that an advanced feature intersection was put in place to mathematically unify different forms of unaligned eye-tracking signals between both COLET & ADABase datasets. The algorithmic logic behind this design can be represented in pseudo-code below:

CL Estimator Pipeline Pseudo Code

```
function PREPROCESS_DATA(raw_data, domain_type)
    if domain is 'COLET' then
        trials ← Extract
    else if domain is 'ADABase' then
        trials ← Segment stream & Mask invalid eyes
    end if
    for each trial in trials do
        features ← Calculate Pupil, Gaze & Blinks
        features ← Normalize Features (Log-Transform)
    end for
    Thresholds ← Calculate quantiles based on NASA-TLX
    labeled_data ← Assign Label (0-3) based on thresholds
    return labeled_data
end function

function ALIGN_DOMAINS(source, target)
    Rename target columns to match source
    Keep only common columns
    Add 'Dataset Source' tag to both
    return aligned_source, aligned_target
end function
```

function RUN_PIPELINE($\mathcal{D}_{source}, \mathcal{D}_{target}$)
$X_{train}, y_{train} \leftarrow \text{StratifiedSplit}(\mathcal{D}_{source}, \text{ratio} = 0.8)$
$X_{balanced}, y_{balanced} \leftarrow \text{RandomOversample}(X_{train}, y_{train})$
$\mathcal{F}_{top\_k} \leftarrow \text{SelectFeaturesRF}(X_{balanced}, y_{balanced}, k = 30)$
$X_{source} \leftarrow \text{FilterFeatures}(X_{balanced}, \mathcal{F}_{top\_k})$
Initial Training
$Model_{Base} \leftarrow \text{TrainHistGradientBoosting}(X_{source}, \mathcal{F}_{top\_k})$
Distribution Alignment (CORAL)
$C_{source} \leftarrow \text{Covariance}(X_{source}) + \lambda I$
$C_{target} \leftarrow \text{Covariance}(X_{target}) + \lambda I$
$C_{align} \leftarrow C_{source}^{-\frac{1}{2}} \cdot C_{target}^{\frac{1}{2}}$
$X_{source}^{aligned} \leftarrow X_{source} \cdot W_{align}$
$Model_{adapted} \leftarrow \text{TrainHistGradientBoosting}(X_{source}^{aligned}, \mathcal{F}_{top\_k})$
Test-Time Adaptation (Self-Training)
$Buffer_{pseudo} \leftarrow \emptyset$
for $Batch_{stream}$ in $\mathcal{D}_{target}^{stream}$ **do**
$Probs \leftarrow Model_{adapted}.\text{PredictProba}(Batch_{stream})$
$Conf_{max} \leftarrow \max(Probs)$
$Label_{pred} \leftarrow \arg\max(Probs)$
if $Conf_{max} > \tau_{high\_confidence}$ **then**
$Buffer_{pseudo} \leftarrow Buffer_{pseudo} \cup \{(Batch_{stream}, Label_{pred})\}$
end if
if $|Buffer_{pseudo}| \geq N_{update\_threshold}$ **then**
Retrain on (Original Source + Pseudo Labeled Target)
$Buffer_{pseudo} \leftarrow \emptyset$
end if
end for
return $Model_{adapted}$
end function

Chapter 5

Result Analysis

5.1 Performance Evaluation

At first, using the preprocessed COLET & ADABase datasets, two separate HistGradientBoostingClassifier models were trained. The goal was to create two individual models which have never seen samples from the other dataset. To validate the baseline scores of these models, we measured their accuracy and both macro & weighted averages of Precision, Recall & F-1 scores. The HistGradientBoostingClassifier is a state-of-the-art ML model, particularly well-suited for tabular data like our pre-processed datasets. This choice was intentional for this four-class classification problem, as per our model specifications defined in section 4.2.1. The baseline HistGradientBoostingClassifier model trained only on the COLET dataset achieved an accuracy score of **78.95%**.

	Precision	**Recall**	**F1-Score**
Macro Average	0.8111	0.7917	0.7927
Weighted Average	0.8123	0.7895	0.7918

Table 5.1: Baseline HistGradientBoostingClassifier on COLET

The other HistGradientBoostingClassifier model trained only on the ADABase dataset achieved a more impressive accuracy score of **84.21%**.

	Precision	**Recall**	**F1-Score**
Macro Average	0.8424	0.8421	0.8398
Weighted Average	0.8424	0.8421	0.8398

Table 5.2: Baseline HistGradientBoostingClassifier on ADABase

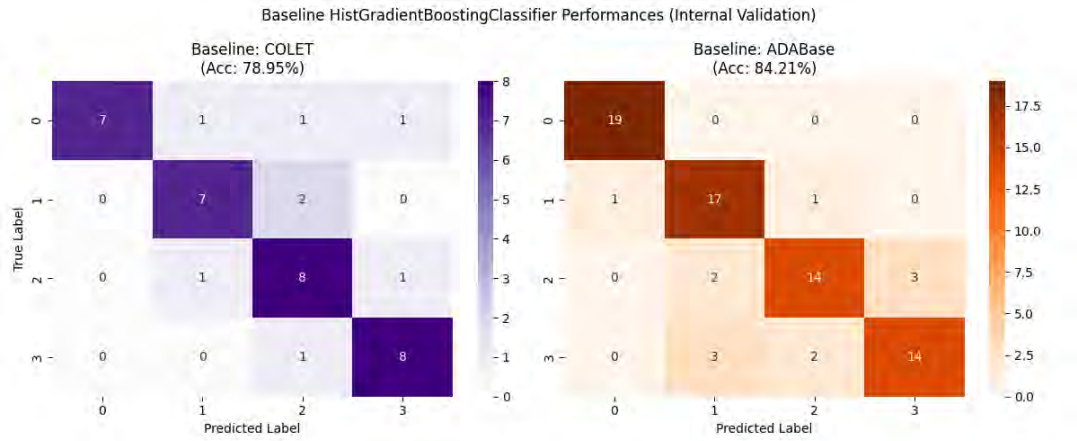

Figure 5.1: Confusion Matrices of the Baseline HistGradientBoostingClassifier

Looking at the Confusion Matrices of these models in Figure 5.1, we can see that both models perform in an acceptable range, with the ADABase model struggling to distinguish between the High (C2) & Very High (C3) classes sometimes, and both models show mostly correct predictions for the Very Low (C0) CL class. This is due to the limited number of observations for classes C2 & C3 in contrast to the abundant C0 entries, as discussed in section 4.3.3.
To validate our choice to use an ML-based pipeline for this tabular data, we also tested the achieved baseline scores against a DL-based state-of-the-art ResNet-MLP model. Despite the theoretical capacity of deep networks, the Residual Multi-Layer Perceptron was only able to achieve an accuracy score of **52.63%** when trained using the COLET dataset.

	Precision	**Recall**	**F1-Score**
Macro Average	0.5366	0.5306	0.5306
Weighted Average	0.5353	0.5263	0.5278

Table 5.3: Baseline ResNet-MLP on COLET

This signifies that our preprocessed tabular data is far too limited in numbers for a DL-based approach to be able to learn all the complex patterns of physiological data, in contrast to EEG-based CL estimation models trained on time series data [43]. This trend is repeated when training the same ResNet-MLP architecture using the ADABase dataset, achieving a similar accuracy score of **53.70%**.

	Precision	**Recall**	**F1-Score**
Macro Average	0.5324	0.5485	0.5388
Weighted Average	0.5295	0.5370	0.5316

Table 5.4: Baseline ResNet-MLP on ADABase

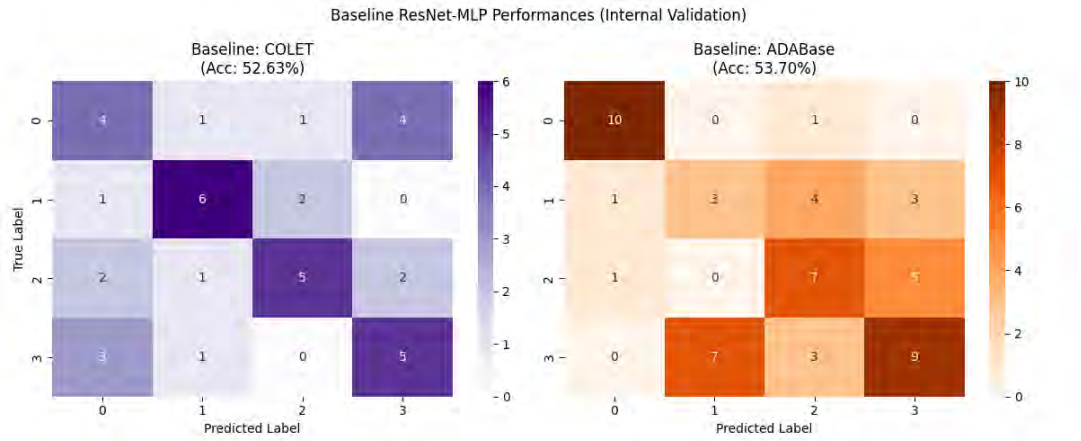

Figure 5.2: Confusion Matrices of the Baseline ResNet-MLP

From the Confusion Matrices of the ResNet-MLP in Figure 5.2, it is evident that in both cases, this architecture struggles to learn the differences between the classes, with the only exception being mostly correct predictions for the C0 class, for the model trained on ADABase only. This indicates underfitting in our DL-based approach, aligning with prior studies on why tree-based models can drastically outperform DL on tabular data [27]. This could be somewhat mitigated by the Synthetic Minority Oversampling Technique (SMOTE); however, that would also change the class distribution in the provided training data and potentially introduce bias, leading to an unreliable model that defeats our purpose.

Now, to analyze the domain gap in each of the selected baseline models (HistGradientBoostingClassifier), an external validation step was performed. Each model was tested on a subset of the unseen target data from the unused dataset. Meaning, on the first pass, the model trained only on COLET was evaluated against a test split randomly taken from ADABase. This yielded a low accuracy score of **22.59%**, which is lower than random guessing.

	Precision	**Recall**	**F1-Score**
Macro Average	0.1779	0.1823	0.1609
Weighted Average	0.1953	0.2259	0.1888

Table 5.5: External Validation on COLET Source -> ADABase Target

On the second pass, the model trained only on ADABase was evaluated against a test split randomly taken from COLET. This also resulted in a poor accuracy score of **21.81%**, with the other performance metrics shown in Table 5.6:

	Precision	**Recall**	**F1-Score**
Macro Average	0.2145	0.2181	0.2097
Weighted Average	0.2145	0.2181	0.2097

Table 5.6: External Validation on ADABase Source -> COLET Target

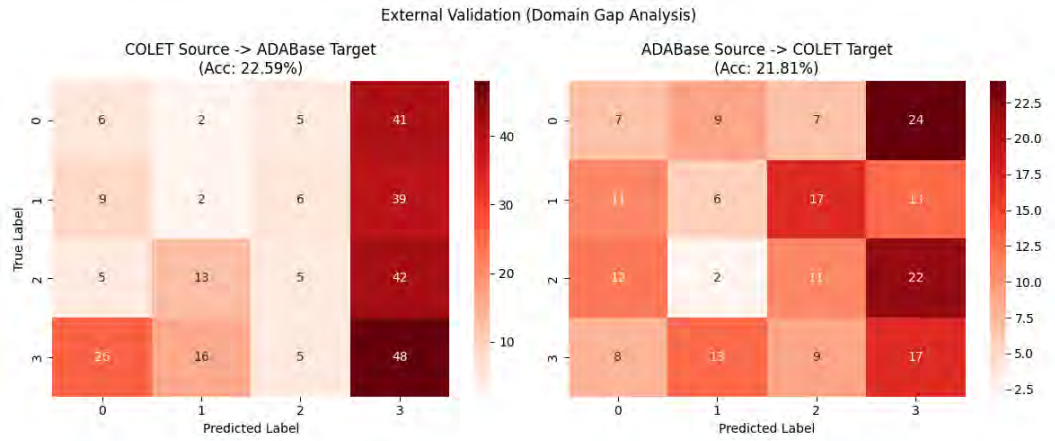

Figure 5.3: Confusion Matrices of the External Validation Tests

From the Confusion Matrices in Figure 5.3, it is clear that on both passes, the external validation tests resulted in a severe loss of accuracy from the baselines (56.36% & 62.4% drops, respectively). This is expected, as previous literature also found CL estimators to be very domain-dependent, as analyzed in section 2.2. This clearly shows a performance gap during domain shifts in physiological data-heavy models like CL estimators, justifying the need for our proposed methodology for cross-domain & test-time adaptation.

5.2 Analysis of Design Solutions

To mitigate the analyzed domain shift, Correlation Alignment (CORAL) was applied. By aligning the second-order statistics of the Source distribution to the Target, an accuracy score of **81.48%** was achieved.

	Precision	**Recall**	**F1-Score**
Macro Average	0.8179	0.8159	0.8141
Weighted Average	0.8174	0.8148	0.8138

Table 5.7: Static CORAL Adaptation

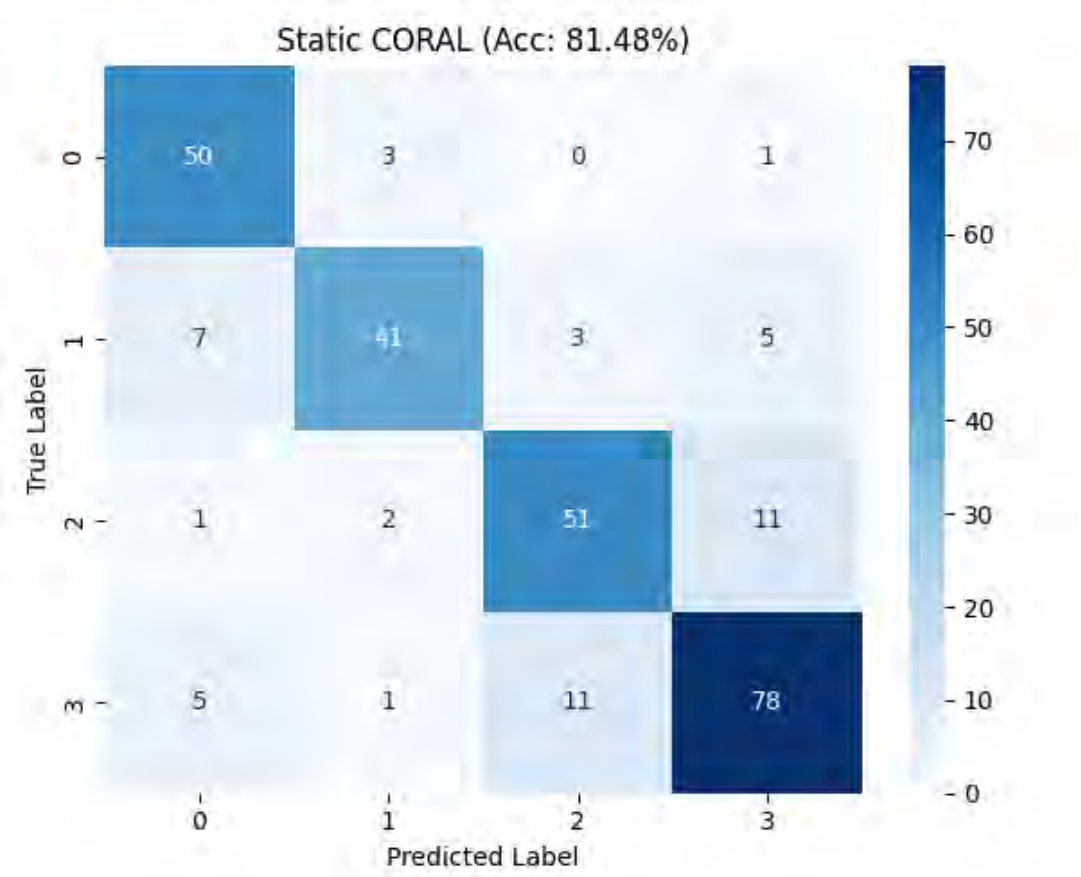

Figure 5.4: Confusion Matrix of the CORAL Adapted Model

Analyzing the Confusion Matrix of the Static CORAL model in Figure 5.4, it is evident that the model is able to make reliable predictions for class C0, with occasional misclassifications between classes C2 & C3 as before.

Now, we have a static model that has a recolored feature space to minimize the covariance distances between the Source & the Target. However, this domain-adapted model still requires ground-truth labels in the Target space and needs to be retrained from scratch if it encounters an unseen target domain. To exemplify, if we introduced another eye-tracking dataset where the target feature space is completely unknown and it may not contain NASA-TLX ground-truth labels, this model would fail. For live, real-world applications, this is a crucial problem.
For this, we implemented iterative pseudo-labeling, such that the model can now produce pseudo-labels based on the prior static, domain-adapted model and a probability distribution generated on an entire unlabeled target dataset. It creates a matrix of probabilities for every sample of the target space, and if the probability is higher than the threshold, it accepts the pseudo-label as an adequate prediction. The high-confidence pseudo-labels are then filtered, and the HistGradientBoosting-Classifier is self-trained on the new, mixed dataset, so that the model has now not only learned from the source logic but also the unseen target structure.
This repeats until the model becomes confident about the previously difficult-to-make predictions (lower confidence) and runs out of new samples to label confidently, denoting that the model has converged.

Thus, the model is able to adjust itself at test-time, pre-inference, without relying on ground-truth labels. This dynamic model achieved an accuracy score of **84.44%**.

	Precision	**Recall**	**F1-Score**
Macro Average	0.8492	0.8417	0.8430
Weighted Average	0.8485	0.8444	0.8442

Table 5.8: Dynamic Test-Time Adaptation

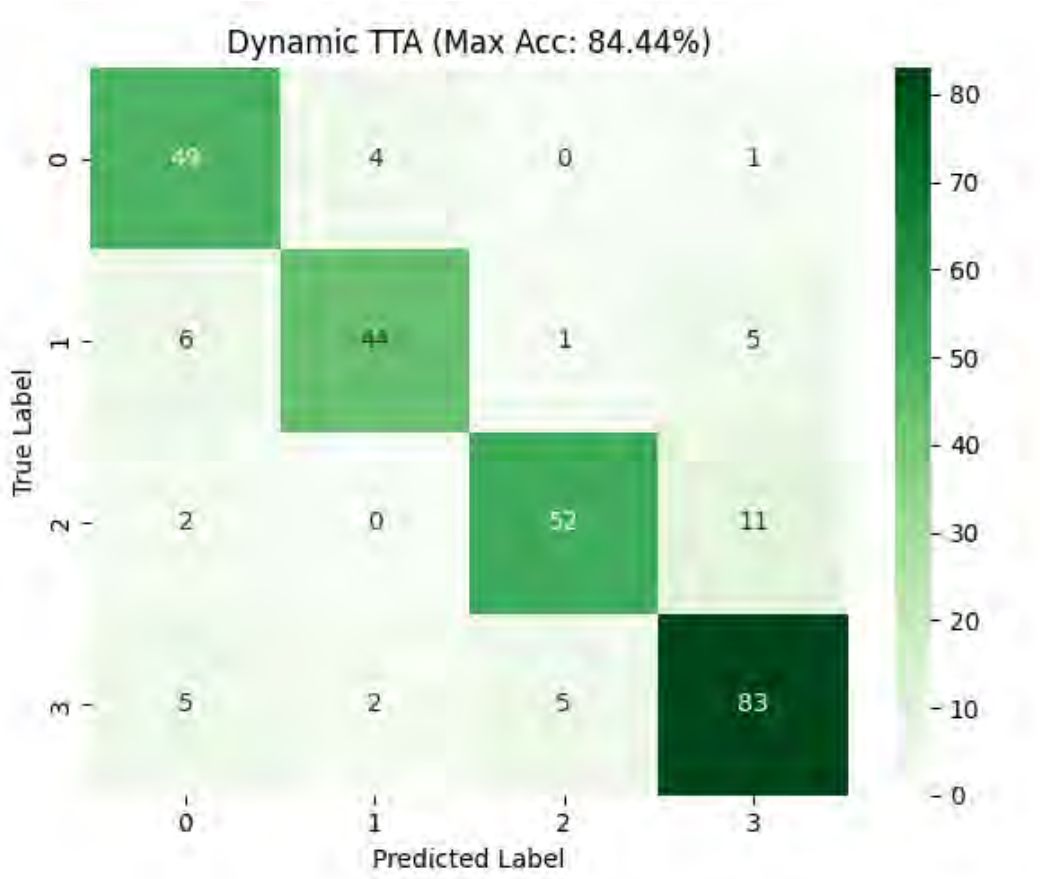

Figure 5.5: Confusion Matrix of the Test-Time Adapted Model

From the Confusion Matrix of the dynamic TTA model in Figure 5.5 & Table 5.8, we can see that this model is not only more suitable for live, unseen data without ground-truth labels, but it also beats our baseline scores & the static domain-adapted model by 2.96%. The model also makes fewer incorrect predictions for the previously challenging C2 & C3 classes, as at runtime, it updates itself by penalizing itself for prior confident, incorrect predictions.

5.3 Final Design Adjustments

For Iterative Pseudo-Labeling, the number of iterations performed can be tweaked to achieve a balance between performance gains (accuracy) & cost (time). Beyond 10 iterations, the law of diminishing returns can be observed, with the log loss increasing over time as well. This denotes that the model becomes overconfident if early stopping conditions are not set. Thus, for real-time applications, a balanced iteration count is set at 10, as the model is able to output predictions within acceptable latency, as shown in Table 5.9, while still outperforming the static, domain-adapted model by a decent margin.

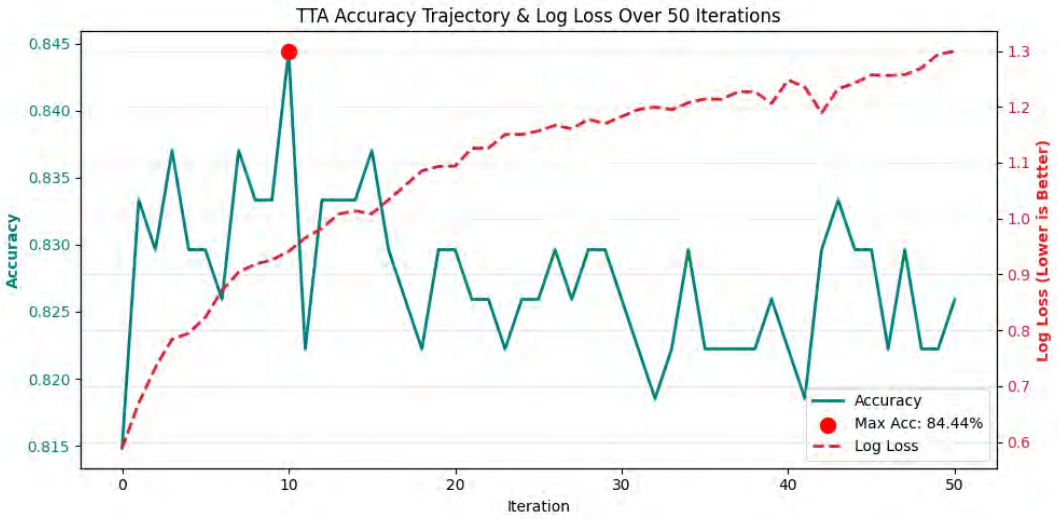

Figure 5.6: TTA Accuracy & Log Loss Trajectory over 50 Iterations

The increasing trend for Log Loss over iterations in Figure 5.6 signifies that the model is being penalized heavily for confident mistakes, despite overall accuracy remaining in a similar range.

Furthermore, to analyze the dynamic TTA model's sensitivity to the confidence threshold, we've run the model back-to-back using a range of confidence thresholds from 0.75 to 0.95, in increments of 0.05. Anything below 75% confidence can be discarded as too unconfident, and anything above 95% confidence is too confident, making the model too rigid.

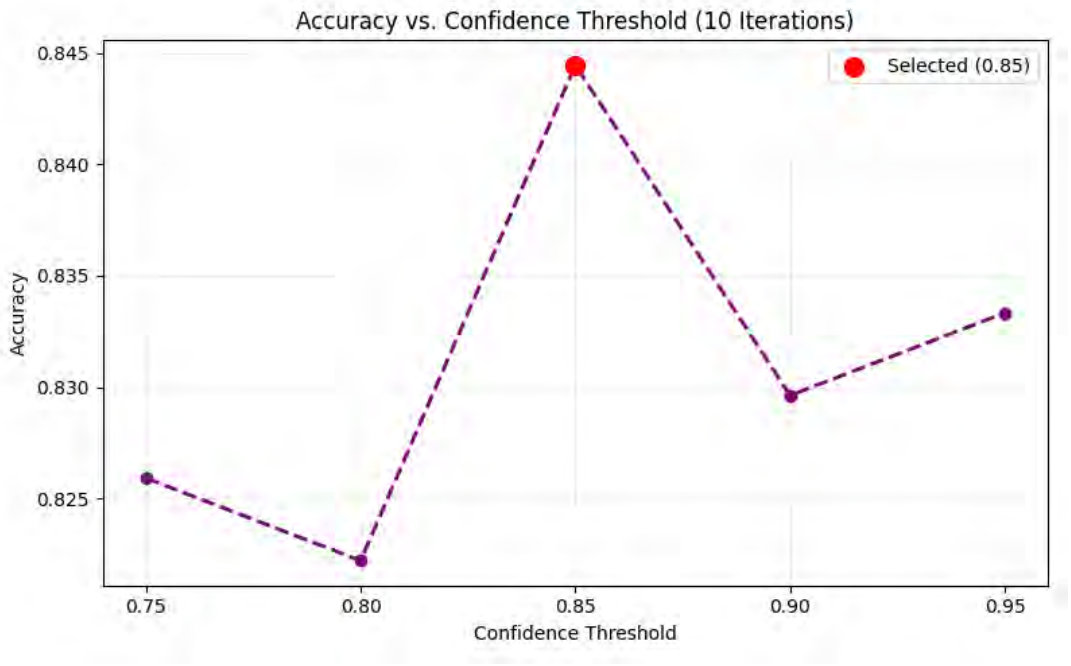

Figure 5.7: TTA Accuracy for varied Confidence Thresholds at 10 iterations

From Figure 5.7, we find that a confidence threshold of 0.85 balances the model from being too flexible or too rigid, which aligns with prior findings in section 2.1 as well [10]. Thus, the final model's confidence threshold was set at 0.85.

5.4 Statistical Analysis

To understand the model's ability to distinguish between the 4 classes, we plotted a multi-class ROC curve and analyzed their separation effect sizes. As the Very High CL class (C3) originally had the fewest samples before class balancing, the model's ability to distinguish between classes C2 & C3 suffered in earlier stages, as seen in the pre-adaptation Confusion Matrices.

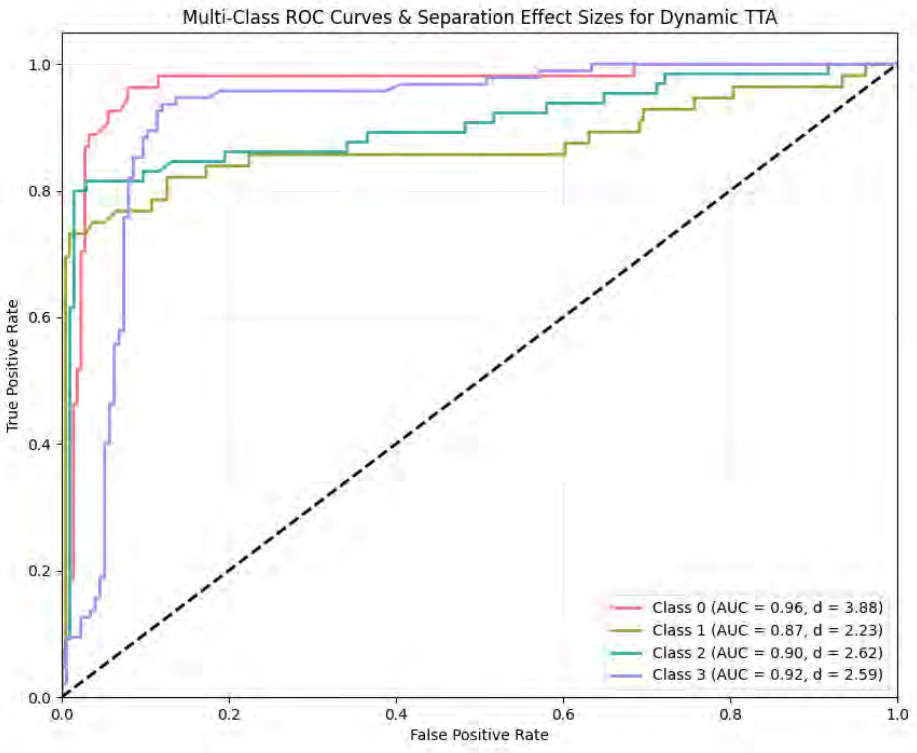

Figure 5.8: Analysis of Cohen's Effect Size on the Dynamic TTA Model

Observing the Cohen's Effect Size coefficients in Figure 5.8, the model can distinguish between all the classes (C0, C1, C2 & C3) with large d values, where $d > 3$ denotes nearly complete separation. Each ROC curve also has a high AUC value, where values above 0.9 signify consistency with the large Cohen's Effect Size coefficients. This entails that our objective of making this CL estimator model compatible with other unified SIID detection frameworks with 4 levels of channel availability (available, slightly affected, affected & unavailable), like HumanI/O, is feasible.
Contrarily, if classes C2 & C3 were to be merged to form a 3-class classification problem of predicting between Low, Medium & High, the overall scores would drastically improve, with the model also being able to distinguish High CL from Medium CL even in the pre-adaptation models. However, since this would also result in improvements across all baseline models, it means that our findings of gradual improvement from the baselines to the static model and, lastly, the dynamic model still hold.

5.5 Comparisons and Relationships

To illustrate the progressive gains from steps 4a to 4d of our proposed methodology, all the obtained accuracy scores have been plotted in a bar chart:

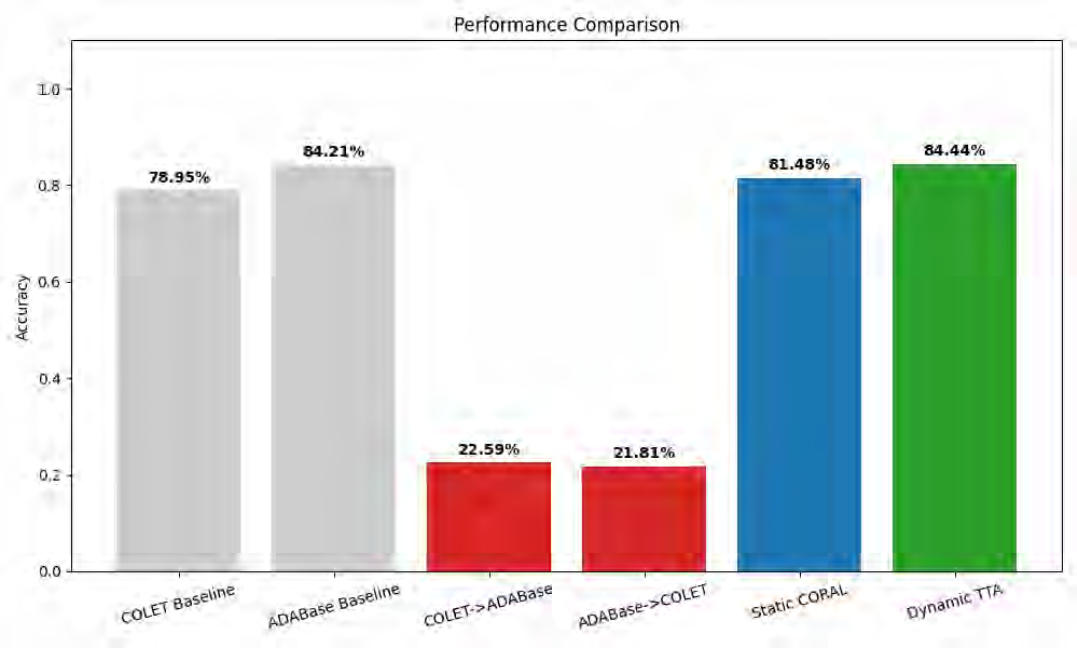

Figure 5.9: Performance Comparison Chart

From Figure 5.9, it is clear that a drastic falloff occurred when the baseline models were externally validated against each other to test for adaptability to domain shifts. This gap was largely made up in the static CORAL model. However, the model still required retraining and ground-truth labels in the target space each time it faced live, unseen data. The dynamic TTA model not only improved upon the static model's performance by iterative self-training, but it is now also adaptable to any new target domain, which was our objective.

To test the latency for live implementations, we ran a performance test using our mid-tier PC from section 3.6, representing a realistic setup to demonstrate the computational efficiency of the dynamic TTA model.

	Accuracy	**Log Loss**	**Batch Time (s)**	**Cumulative Time (s)**
Iteration 10	0.8444	0.9409	0.5634	5.0578
Iteration 20	0.8296	1.0940	0.6928	11.3697
Iteration 30	0.8259	1.1825	0.6959	18.1747
Iteration 40	0.8222	1.2480	0.8140	25.7121
Iteration 50	0.8259	1.2996	0.8474	33.7822

Table 5.9: TTA Performance Evaluation

As the HistGradientBoostingClassifier from scikit-learn does not natively support CUDA and runs only on the CPU, this represents a reproducible scenario on many consumer-grade processors with integrated GPUs.

Table 5.9 shows that our early stopping point of 10 iterations took just 5 seconds to run on CPU. As the iteration count increased, the batch processing time also increased rapidly, indicating that the longer the iterations continued, the more time it took per batch. Thus, the time delta between any two iterations is not fixed. The first iteration had the lowest batch time, at 0.4659 seconds, while the last (50th) iteration had the highest, at 0.8474 seconds. This is why running 50 iterations took 33.78 seconds to complete instead of a value closer to 25 seconds.

From this comparison, we can conclude that this dynamic TTA model is lightweight enough for latency-sensitive applications. While 10 iterations needing 5 seconds is still beyond the margin of an acceptable range for real-time retraining and inference, it is vastly more efficient than DL-based approaches like DANN & TENT adaptation. Despite the computational complexity, DANN yields an accuracy score of **75.45%** on EEG-based time series data [43], scoring just below the tabular data-based static CORAL adapted model. The dynamic TTA model's performance can be further optimized when paired with higher-end CPUs, smaller target data, or by choosing an even lower iteration count. To exemplify, at iteration 3, the accuracy score already reached **83.70%**, and it took only 1.4840 seconds to complete. However, this risks proper adaptation to the target domain, as there could still be more unconfident pseudo-labels left.

5.6 Discussions

5.6.1 Interpretation of the Results via XAI

We have extracted the top 10 features based on the mean accuracy decrease across all trained models for explainability. To rank them by importance, the importance vectors were computed and validated using permutation. Permutation importance scores are model-agnostic and less biased towards high-cardinality features, which is why we chose to extract each of the feature importance charts using this method.

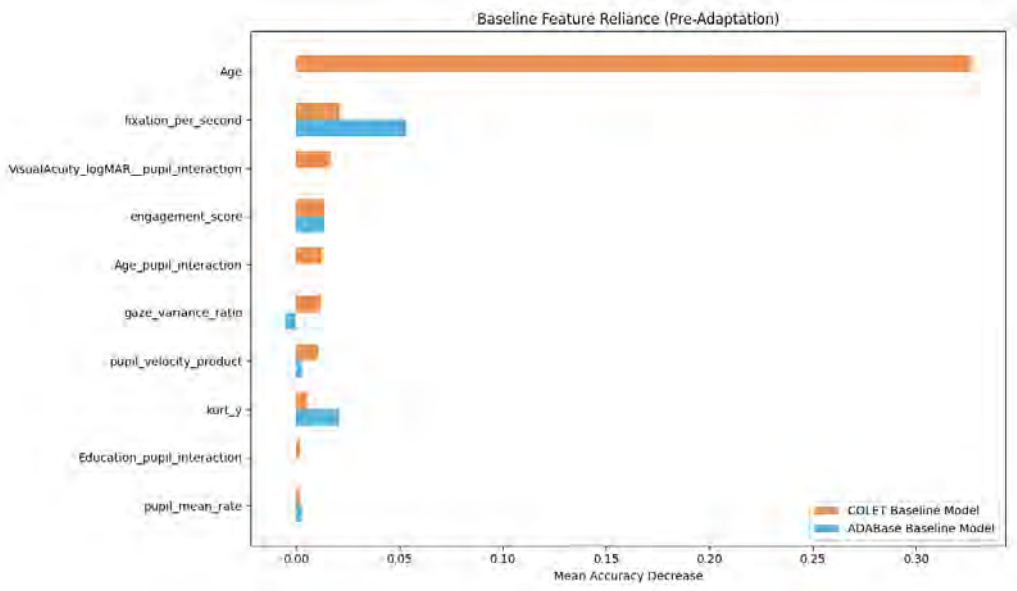

Figure 5.10: Feature Importance of the Baseline Models before Adaptation

Figure 5.10 shows the baseline models' feature importance chart, sorted by the rankings of the HistGradientBoostingClassifier model trained on COLET. We can see that the subject Age accounts for the largest mean accuracy decrease in the baseline COLET-trained model, whereas this feature has nearly 0 permutation importance in the ADABase-trained model. Similarly, the rest of the features either have dissimilar feature importances, 0 permutation importance, or even negative permutation importance scores, particularly in the gaze variance rates.
This explains the drastic performance degradation when the external validation step, 4b, was performed, as seen in Figure 5.9, since the source & target have vastly different feature importances. Moreover, Age is a subjective metadata. It may or may not be present in other eye-tracker datasets. While age may have a correlation with perceived task load (NASA-TLX), when domain shifts occur, if the model prioritizes this feature and the inference is based on the subject's age, it could be a potentially biased prediction.

This is exactly the problem Correlation Alignment tries to solve. In our static CORAL model, we can observe that the source & target domains have been adapted and the subjective features have been deprioritized:

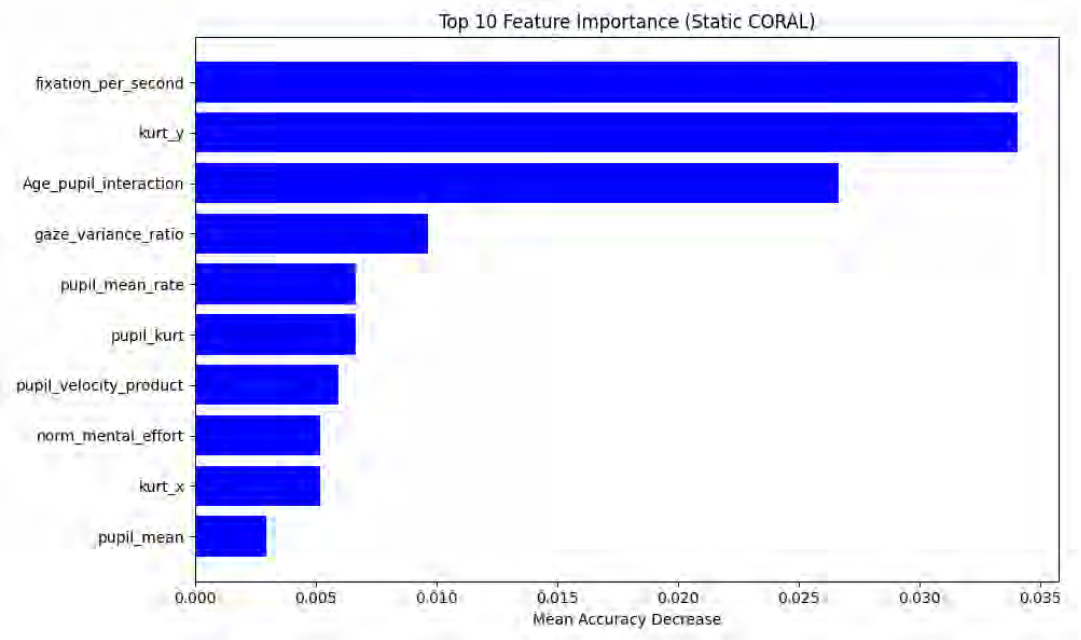

Figure 5.11: Feature Importance of the Static Model after CORAL Adaptation

Figure 5.11 shows that the most important feature in the static CORAL adapted model is the fixations per second metric, which, according to Cognitive Load Theory, should be a prioritized feature in estimating CL from eye-tracking. This is followed by the second most important feature, kurtosis of the gaze in the y-axis. Kurtosis is a measure of the peakedness of a probability distribution. A high kurtosis with heavy tails (outliers) signifies that the user is fixing their gaze mostly on a specific vertical level, but occasionally makes large vertical jumps (saccades). This is also a key indicator of attentional tunneling, which happens under high cognitive load. On the contrary, a lower kurtosis entails that the user is scanning vertically, in a more regular, distributed pattern.

The pupil interaction scores from the engineered features are also prevalent in this ordered list, which aligns with our findings in section 2.2, where pupil dilation was deemed an effective indicator of CL [7].
The notable takeaway is that the topmost important feature's mean accuracy decrease has reduced to a tenth of the baseline model's most important feature, whilst prioritizing domain-invariant features correlated to the label. This implies that CORAL was able to adapt the source & target domains adequately, which resulted in the performance gains.

To interpret what the dynamic TTA model prioritized, we similarly extracted the top 10 feature importance chart for the TTA model with 10 iterations and a confidence threshold of 0.85:

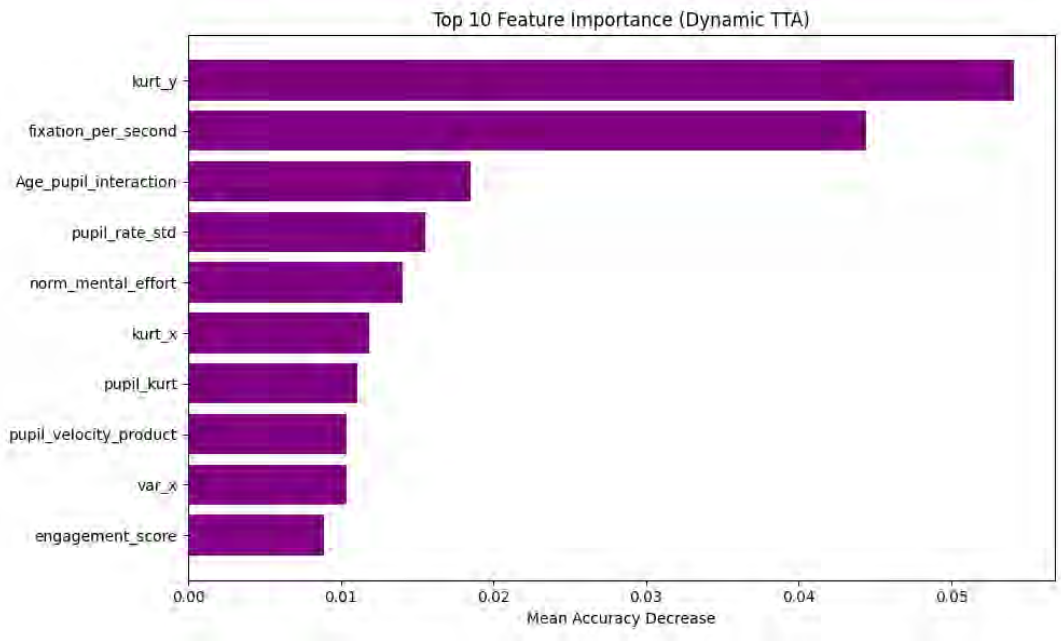

Figure 5.12: Feature Importance of the Dynamic Model after Test-Time Adaptation

As observed from Figure 5.12, TTA seems to prioritize kurtosis of the gaze in the y-axis even further, switching places with fixations per second. The other feature importances remain comparable as before, with the feature-engineered pupil interaction scores remaining in the topmost prioritized list. Although the maximum mean accuracy decrease value increases slightly from the static CORAL adapted model, as depicted by the values on the x-axis of the plot in Figure 5.12, this is still far below the baseline models' maximum. This means that, during iterative pseudo-labeling, the model had to update itself to prioritize values of kurtosis of the gaze even more, in order to avoid being penalized for making incorrect predictions with high confidence.

While the permutation scores highlight how the performance would drop if a particular feature were to be removed, SHAP analysis complements this finding by allowing an insight into which features contribute the most to the decision-making process of the model. Analyzing the global feature importance values retrieved from SHAP:

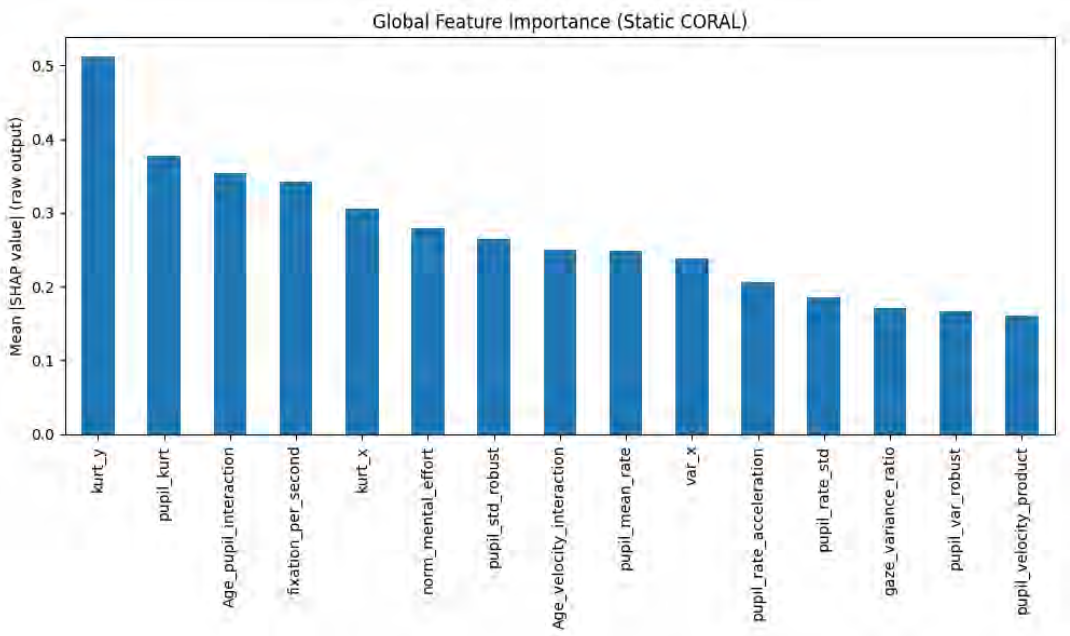

Figure 5.13: Global Feature Importance of the Static CORAL Model

From Figure 5.13, it appears that kurtosis in the y-axis is actually the top contributing feature in the static model's predictions too, aligning with our outcomes from the feature importance chart of the Test-Time Adapted model, with fixations per second having a lower rank.

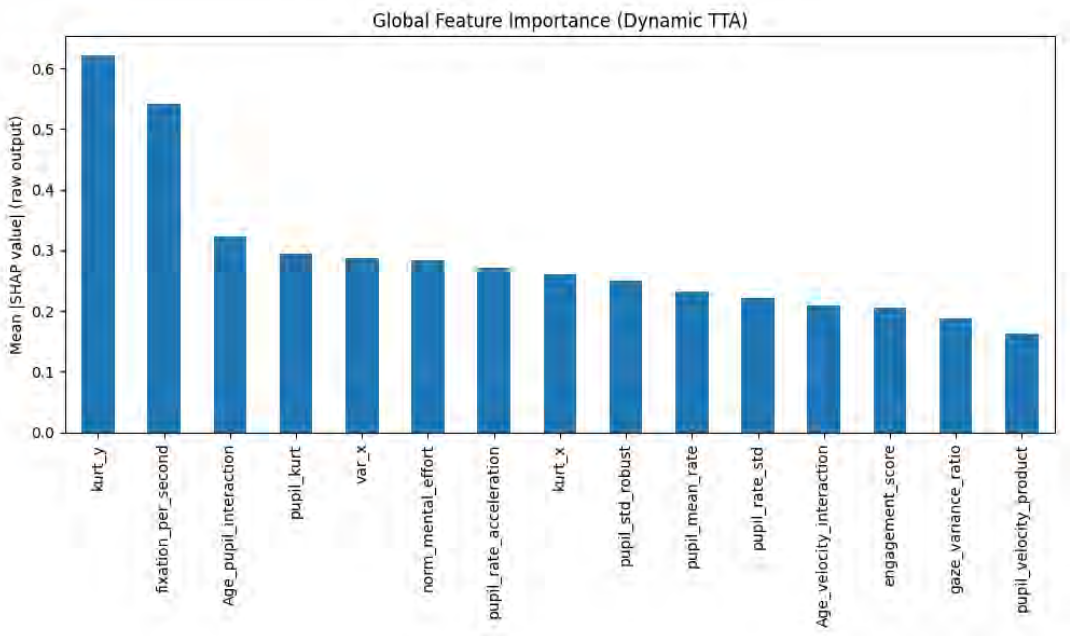

Figure 5.14: Global Feature Importance of the Dynamic TTA Model

After Test-Time Adaptation, in Figure 5.14, it appears that the fixations per second feature gained priority in decision-making, ranked at the second most important feature, with kurtosis in the y-axis having an even higher mean SHAP value, which aligns with the local feature importance.

To analyze which groups of physiological eye features contributed the most in the TTA-adapted model's decision chain, we calculated total mean SHAP values for 3 groups - Pupil, Fixation & Blink, where each individual feature's SHAP values were appended to its respective group via simple membership operations. Beyond the physiological features, all other behavioral features, particularly the features obtained from our feature engineering chain to calculate interaction scores, were denoted as higher-order gaze behavior.

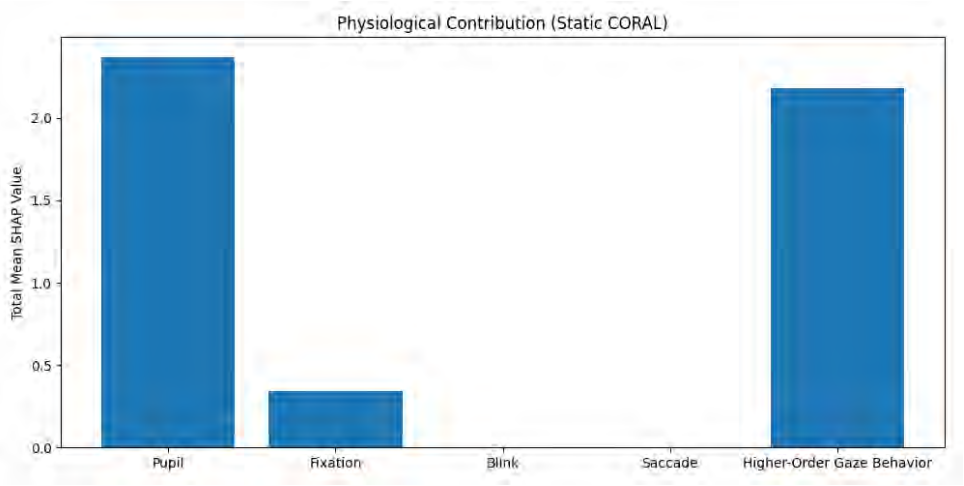

Figure 5.15: Contributions of Physiological Groups of the Static CORAL Model

From Figure 5.15, we can see that the Pupil group contributes the most, with the highest total mean SHAP value. This calls back our findings from past literature, denoting Pupil Dilation as an effective indicator of CL [7]. On the contrary, fixation had a low contribution, and blink & saccade metrics had zero contribution. This is because the blink data in the source distribution was sparse and often task-dependent and subject-specific, as seen in section 4.3.1. On the contrary, saccades are high-frequency but short-lived and often recorded indirectly as a fixation-derived metric, making them harder to capture. Thus, the model is actually not relying on these second-order features that are less generalizable, which is a net positive factor for the model.

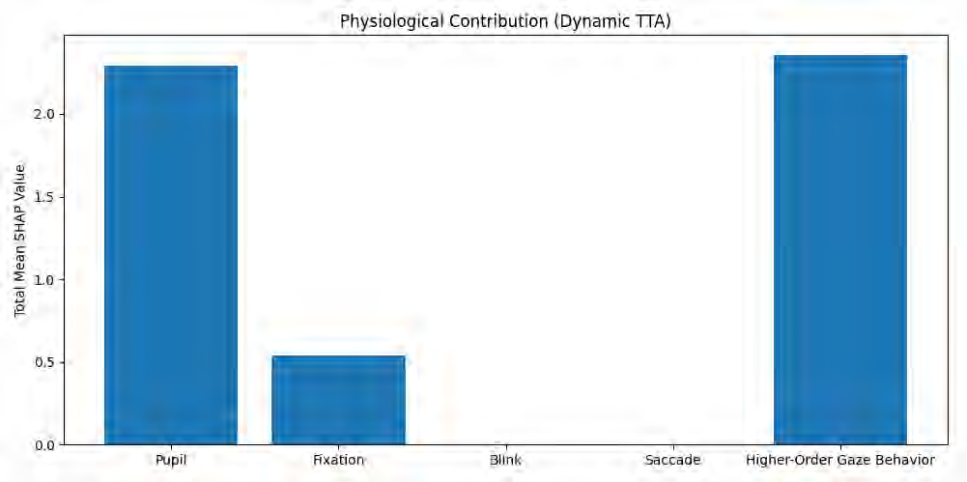

Figure 5.16: Contributions of Physiological Groups of the Dynamic TTA Model

As for the dynamic TTA model, from Figure 5.16, we can see that the grouped contributions mostly remain similar, with a heightened total mean SHAP value in the higher-order gaze behavior. This denotes that the model had to readjust itself to rely more on the behavioral features derived from our feature engineering pipeline, beyond just the physiological groups. As for the physiological contributions, it accurately prioritizes the key estimator of CL - the pupil dilation and pupil-related metrics. This makes the model less subjective, signifying adequate adaptation to the target domain.

Lastly, to analyze SHAP rank stability before & after Test-Time Adaptation, the Spearman rank correlations were calculated. A high rank correlation ($\rho > 0.75$) generally demonstrates strong consistency.

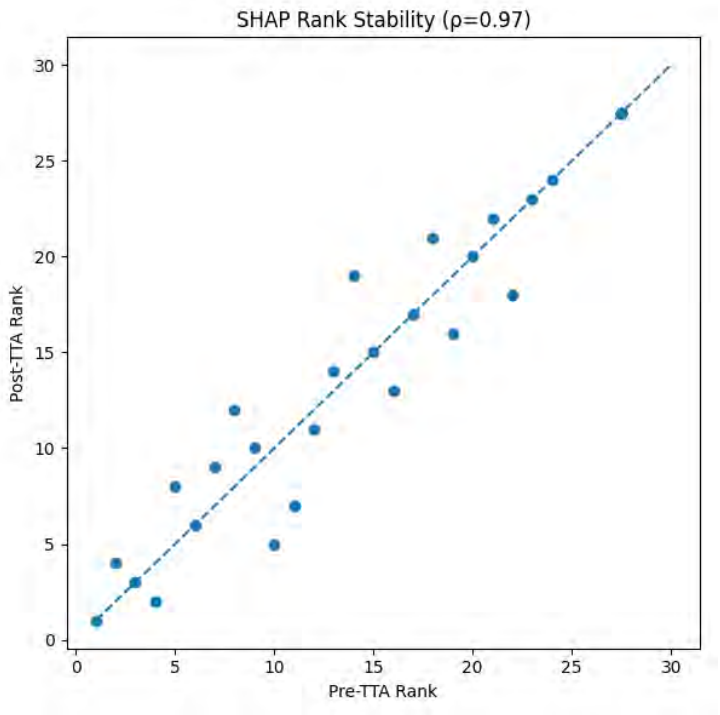

Figure 5.17: SHAP Rank Stability before & after TTA

We obtained a strong rank correlation of $\rho = 0.97$. From Figure 5.17, it is evident that the model does not distort the feature reliance in the process of adapting at test-time. This confirms that the learned representation of CL remains stable across domains, resulting in a dynamic model that was not only aptly domain-adapted but can also make reliable predictions on live, unseen target data without ground-truth labels, leading to less overall bias during inference.

5.6.2 Limitations

While TTA does make the model dynamic, implementations of this model would still require the live input data to undergo the same pre-processing, feature engineering & feature alignment steps to obtain similar results. While this remains calibration-free, it denotes shortcomings in the real-world applications of this CL estimator model. Depending on the eye-tracker device in use, the format of the tabular data may differ, as there is no unified approach or standard for recording this data. As observed in section 4.3, COLET & ADABase had to go through varied pre-processing chains to yield an aligned feature space. This signifies the need for standardization of eye-tracker data streams, which is a limitation of this model.

Another limitation is the fact that HistGradientBoostingClassifier is an ensemble model based on decision trees. This means that, unlike DL-based models, where semi-supervised learning pipelines can be incrementally updated via gradient descent, the entire decision tree needs to be rebuilt each time the dynamic TTA model updates itself. This can be computationally demanding on rather complex decision trees; however, in our tabular data, that was not the case, as analyzed in section 5.5 (refer to Table 5.9).

Finally, our decision not to merge classes C2 & C3 for a 3-class classification problem means that the model's peak performance is hindered. However, our objective was to demonstrate the model's adequacy in adapting to the Target domain, and compatibility with existing SIID detection frameworks, not raw accuracy scores in benchmarks. As a traditional 3-class classification would also improve random guessing accuracy & all the baseline scores, our findings of gradual improvements post-adaptation remain.

Chapter 6

Conclusion

6.1 Summary of Findings

To summarize, our findings have aligned with past literature in validating the need for a lightweight, calibration-free adaptation pipeline. The efficacy of Correlation Alignment yields remarkable gains over the models before adaptation. While this is an improvement, it still creates a static model that requires target labels and would have to be retrained manually in case of adapting to a new target. Iterative-Pseudo Labeling creates a dynamic, self-supervised test-time adapted model that not only gains in accuracy but also does not require any ground-truth labels. Our findings show that this dynamic model is able to retrain itself to prioritize domain-independent features relevant to CL estimation in making a less biased prediction than the pre-adaptation models. Interpreting the results of the model shows how it deprioritized subjective metrics to make the predictions more neutral. This implementation remains lightweight enough to run within acceptable margins of latency on a variety of end devices, in contrast to more computationally demanding DL-based approaches.
As we step into the future where wearable XR devices integrate with our daily lives, dynamic CL classifiers that can adapt to diverse environments in real time present a huge potential in aiding with accessibility within the future of interactive systems.

6.2 Contributions to the Field

As the dynamic model does not require ground truth labels and outperforms not only the static model but also the baselines, it signifies that our objective to create a dynamic model that adapts to new data in real time was met, whilst offering validation for the obtained results. We have shown that:

- The Domain Shift Gap is present even when a Source dataset has been methodically preprocessed & the Target dataset has been feature aligned with the Source.
- Correlation Alignment is a suitable Domain Adaptation strategy on smaller, tabular datasets, as an alternative to Domain-Adversarial Neural Networks, which requires significantly more training data to perform adequately.

- The computational efficiency of Iterative Pseudo-Labeling as a Test-Time Adaptation strategy is lightweight enough as a self-supervised model to run on consumer-grade, CPU-only devices, whilst offering the most reliable predictions in our testing, without requiring manual interventions.
- By setting a balanced confidence threshold & an early stopping point, the dynamic TTA model avoids becoming overconfident and makes reliable predictions without requiring explicit ground-truth labels.

6.3 Recommendations for Future Work

For future work, avenues that could be explored include, but are not limited to:

- Integrating multimodal sensor fusion and comparing the effectiveness of time series EEG data against tabular eye-tracker data for CL estimation
- Implementing a Deep Learning-based pipeline for Cross-Domain & Test-Time Adaptation that works well on both tabular data & time series data
- Comparing lightweight ML-based ensemble models to deep neural network-based architectures in their efficacy of proper domain adaptation techniques via feature reliance analysis
- Developing a signal chain that undergoes minimal preprocessing to produce validated, reliable predictions to make a truly schema-agnostic, semantically interoperable CL estimator model
- Creating a standardized format for sampling user gaze information, such that shifting targets do not require prior feature alignment
- Comparing the dependability of level-based classification (i.e., 4-classes, 3-classes) to gradients of availability (probability distribution) in estimating CL
- Developing a privacy-preserving framework for extracting gaze data
- Pipelining an adapted CL estimator model to existing, unified SIID detection frameworks, for proper multimodal detection of a user's channel availability that includes their cognitive states

Bibliography

[1] D. Kahneman, *Attention and Effort*. Englewood Cliffs, NJ: Prentice-Hall, 1973.

[2] J. Beatty, "Task-evoked pupillary responses, processing load, and the structure of processing resources," *Psychological Bulletin*, vol. 91, no. 2, pp. 276–292, 1982. DOI: 10.1037/0033-2909.91.2.276.

[3] S. G. Hart and L. E. Staveland, "Development of nasa-tlx (task load index): Results of empirical and theoretical research," in *Human Mental Workload*, ser. Advances in Psychology, P. A. Hancock and N. Meshkati, Eds., vol. 52, North-Holland, 1988, pp. 139–183. DOI: https://doi.org/10.1016/S0166-4115(08)62386-9. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0166411508623869.

[4] E. Granholm, R. F. Asarnow, A. J. Sarkin, and K. L. Dykes, "Pupillary responses index cognitive resource limitations," *Psychophysiology*, vol. 33, no. 4, pp. 457–461, 1996. DOI: 10.1111/j.1469-8986.1996.tb01071.x.

[5] S. Marshall, "The index of cognitive activity: Measuring cognitive workload," in *Proceedings of the IEEE 7th Conference on Human Factors and Power Plants*, 2002, pp. 7–7. DOI: 10.1109/HFPP.2002.1042860.

[6] A. Gevins and M. E. Smith, "Neurophysiological measures of cognitive workload during human-computer interaction," *Theoretical Issues in Ergonomics Science*, vol. 4, no. 1-2, pp. 113–131, 2003. DOI: 10.1080/14639220210159717. eprint: https://doi.org/10.1080/14639220210159717. [Online]. Available: https://doi.org/10.1080/14639220210159717.

[7] M. Pomplun and S. Sunkara, "Pupil dilation as an indicator of cognitive workload in human-computer interaction," 2003. [Online]. Available: https://api.semanticscholar.org/CorpusID:1052200.

[8] F. Jiao and G. He, "Real-time eye detection and tracking under various light conditions," *Data Science Journal*, Oct. 2007. DOI: 10.2481/dsj.6.S636.

[9] S.-R. Ke, H. L. U. Thuc, Y.-J. Lee, J.-N. Hwang, J.-H. Yoo, and K.-H. Choi, "A review on video-based human activity recognition," *Computers*, vol. 2, no. 2, pp. 88–131, 2013, ISSN: 2073-431X. DOI: 10.3390/computers2020088. [Online]. Available: https://www.mdpi.com/2073-431X/2/2/88.

[10] D.-H. Lee, "Pseudo-label : The simple and efficient semi-supervised learning method for deep neural networks," 2013. [Online]. Available: https://api.semanticscholar.org/CorpusID:18507866.

[11] H. J. Oh, H. C. Kim, H. Hwangbo, and Y. G. Ji, "User centered inclusive design process: A 'situationally-induced impairments and disabilities' perspective," in *Human-Computer Interaction. Human-Centred Design Approaches, Methods, Tools, and Environments*, M. Kurosu, Ed., Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 103–108, ISBN: 978-3-642-39232-0.

[12] K. Stojmenova and J. Sodnik, "Methods for assessment of cognitive workload in driving tasks," in *Proceedings of the International Conference on Information Society*, 2015, pp. 229–234. [Online]. Available: http://www.eventiotic.com/eventiotic/library/paper/123.

[13] T. R. Hayes and A. A. Petrov, "Mapping and correcting the influence of gaze position on pupil size measurements," *Behavior Research Methods*, vol. 48, no. 2, pp. 510–527, 2016. DOI: 10.3758/s13428-015-0588-x.

[14] B. Sun, J. Feng, and K. Saenko, "Correlation alignment for unsupervised domain adaptation," in *Domain Adaptation in Computer Vision Applications*, ser. Advances in Computer Vision and Pattern Recognition, 2017, pp. 153–171. DOI: 10.1007/978-3-319-58347-1_8. [Online]. Available: https://doi.org/10.1007/978-3-319-58347-1_8.

[15] A. Gustafsson, *Eye movement analysis for activity recognition in everyday situations*, 2018.

[16] M. C. Schall, R. F. Sesek, and L. A. Cavuoto, "Barriers to the adoption of wearable sensors in the workplace: A survey of occupational safety and health professionals," *Human Factors*, vol. 60, no. 3, pp. 351–362, 2018. DOI: 10.1177/0018720817753907. [Online]. Available: https://doi.org/10.1177/0018720817753907.

[17] M. M. Baig, H. GholamHosseini, S. Afifi, and F. Mirza, "Current challenges and barriers to the wider adoption of wearable sensor applications and internet-of-things in health and well-being," in *Proceedings of the CONF-IRM 2019 International Conference on Information Resources Management*, 2019, p. 20. [Online]. Available: https://aisel.aisnet.org/confirm2019/20.

[18] E. Kazakos, A. Nagrani, A. Zisserman, and D. Damen, "Epic-fusion: Audio-visual temporal binding for egocentric action recognition," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 5492–5501.

[19] F. Göbel, K. Kurzhals, M. Raubal, and V. R. Schinazi, *Gaze-aware mixed-reality: Addressing privacy issues with eye tracking*, ETH Zurich Repository for Publications and Research Data, 2020. DOI: 10.3929/ethz-b-000409514. [Online]. Available: https://doi.org/10.3929/ethz-b-000409514.

[20] Y. G. Kim and C. Wu, "Autoscale: Energy efficiency optimization for stochastic edge inference using reinforcement learning," in *Proceedings of the IEEE/ACM International Symposium on Software Engineering for Adaptive and Self-Managing Systems (SEAMS)*, 2020, pp. 1082–1096. DOI: 10.1109/MICRO50266.2020.00090. [Online]. Available: https://doi.org/10.1109/MICRO50266.2020.00090.

[21] E. P. Larsen, J. M. Kolman, F. N. Masud, and F. Sasangohar, "Ethical considerations when using a mobile eye tracker in a patient-facing area: Lessons from an intensive care unit observational protocol," *Ethics & Human Research*, vol. 42, no. 6, pp. 2–13, 2020. DOI: 10.1002/eahr.500068. [Online]. Available: https://doi.org/10.1002/eahr.500068.

[22] T. Appel, *Cross-participant and Cross-task Classification of Cognitive Load Based on Eye Tracking*. Eberhard Karls Universität Tübingen, 2021. [Online]. Available: https://books.google.com.bd/books?id=etF90AEACAAJ.

[23] R. E. Bixler and S. K. D'Mello, "Crossed eyes: Domain adaptation for gaze-based mind wandering models," in *Proceedings of the ACM Symposium on Eye Tracking Research and Applications (ETRA)*, 2021, pp. 1–12. DOI: 10.1145/3448017.3457386. [Online]. Available: https://doi.org/10.1145/3448017.3457386.

[24] M. Kaczorowska, P. Karczmarek, M. Plechawska-Wójcik, and M. Tokovarov, "On the improvement of eye tracking-based cognitive workload estimation using aggregation functions," *Sensors*, vol. 21, no. 13, 2021, ISSN: 1424-8220. DOI: 10.3390/s21134542. [Online]. Available: https://www.mdpi.com/1424-8220/21/13/4542.

[25] J. Liang, D. Hu, and J. Feng, *Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation*, 2021. arXiv: 2002.08546 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2002.08546.

[26] A. Aygun, B. Lyu, T. Nguyen, Z. Haga, S. Aeron, and M. Scheutz, "Cognitive workload assessment via eye gaze and eeg in an interactive multi-modal driving task," in *Proceedings of the 2022 International Conference on Multimodal Interaction*, ser. ICMI '22, Bengaluru, India: Association for Computing Machinery, 2022, pp. 337–348, ISBN: 9781450393904. DOI: 10.1145/3536221.3556610. [Online]. Available: https://doi.org/10.1145/3536221.3556610.

[27] L. Grinsztajn, E. Oyallon, and G. Varoquaux, "Why do tree-based models still outperform deep learning on typical tabular data?" *Advances in neural information processing systems*, vol. 35, pp. 507–520, 2022.

[28] E. Ktistakis, V. Skaramagkas, D. Manousos, *et al.*, "Colet: A dataset for cognitive workload estimation based on eye-tracking," *Computer Methods and Programs in Biomedicine*, vol. 224, p. 106 989, 2022.

[29] M. P. Oppelt, A. Foltyn, J. Deuschel, *et al.*, "Adabase: A multimodal dataset for cognitive load estimation," *Sensors*, vol. 23, no. 1, p. 340, 2022.

[30] F. N. Biondi, F. Graf, P. Pillai, and B. Balasingam, "On validating a generic camera-based blink detection system for cognitive load assessment," *Cognitive Computation and Systems*, vol. 5, no. 4, pp. 255–264, 2023. DOI: https://doi.org/10.1049/ccs2.12088. eprint: https://ietresearch.onlinelibrary.wiley.com/doi/pdf/10.1049/ccs2.12088. [Online]. Available: https://ietresearch.onlinelibrary.wiley.com/doi/abs/10.1049/ccs2.12088.

[31] E. Iacobelli, V. Ponzi, S. Russo, and C. Napoli, "Eye-tracking system with low-end hardware: Development and evaluation," *Information*, vol. 14, no. 12, 2023, ISSN: 2078-2489. DOI: 10.3390/info14120644. [Online]. Available: https://www.mdpi.com/2078-2489/14/12/644.

[32] A. Korkmaz and S. Gülseçen, "The intersection of ethics and biometric data in eye-tracking systems," in *Innovative Research in Social, Human and Administrative Sciences*, 2023, p. 528. DOI: 10.59287/irshas.402. [Online]. Available: https://doi.org/10.59287/irshas.402.

[33] A. Pouliou, F. Kehagia, G. Poulios, M. Pitsiava-Latinopoulou, and E. Bekiaris, "Drivers' reaction time and mental workload: A driving simulation study," *Transport and Telecommunication Journal*, vol. 24, no. 4, pp. 397–408, 2023. DOI: 10.2478/ttj-2023-0031. [Online]. Available: https://doi.org/10.2478/ttj-2023-0031.

[34] M. C. Sáiz-Manzanares, R. Marticorena-Sánchez, L. J. Martín-Antón, L. Almeida, and M. A. Carbonero-Martín, "Application and challenges of eye tracking technology in higher education," *Comunicar*, vol. 31, no. 76, 2023. DOI: 10.3916/C76-2023-03. [Online]. Available: https://doi.org/10.3916/C76-2023-03.

[35] N. M. González and E. Bozkir, "Eye-tracking devices for virtual and augmented reality metaverse environments and their compatibility with the european union general data protection regulation," *Digital Society*, vol. 3, no. 2, 2024. DOI: 10.1007/s44206-024-00128-9. [Online]. Available: https://doi.org/10.1007/s44206-024-00128-9.

[36] X. B. Liu, J. N. Li, D. Kim, X. '. Chen, and R. Du, "Human i/o: Towards a unified approach to detecting situational impairments," in *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, ser. CHI '24, Honolulu, HI, USA: Association for Computing Machinery, 2024, ISBN: 9798400703300. DOI: 10.1145/3613904.3642065. [Online]. Available: https://doi.org/10.1145/3613904.3642065.

[37] J. Sweller, "Cognitive load theory and individual differences," *Learning and Individual Differences*, vol. 110, p. 102 423, 2024, ISSN: 1041-6080. DOI: https://doi.org/10.1016/j.lindif.2024.102423. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1041608024000165.

[38] A. Tabbakh, L. A. Amin, M. Islam, G. M. I. Mahmud, I. K. Chowdhury, and M. S. H. Mukta, "Towards sustainable ai: A comprehensive framework for green ai," *Discover Sustainability*, vol. 5, no. 1, 2024. DOI: 10.1007/s43621-024-00641-4. [Online]. Available: https://doi.org/10.1007/s43621-024-00641-4.

[39] Z. Xiao and C. G. M. Snoek, "Beyond model adaptation at test time: A survey," *arXiv*, 2024. DOI: 10.48550/arXiv.2411.03687. [Online]. Available: https://doi.org/10.48550/arXiv.2411.03687.

[40] Y. Xu, H. Cao, K. Mao, Z. Chen, L. Xie, and J. Yang, "Aligning correlation information for domain adaptation in action recognition," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 35, no. 5, pp. 6767–6778, 2024. DOI: 10.1109/TNNLS.2022.3212909.

[41] E. Dritsas, M. Trigka, C. Troussas, and P. Mylonas, "Multimodal interaction, interfaces, and communication: A survey," *Multimodal Technologies and Interaction*, vol. 9, no. 1, 2025, ISSN: 2414-4088. DOI: 10.3390/mti9010006. [Online]. Available: https://www.mdpi.com/2414-4088/9/1/6.

[42] A. Rehman, A. Dæhlen, I. Heldal, and J. C. Lin, "Privacy preservation and identity tracing prevention in ai-driven eye tracking for interactive learning environments," *IEEE Access*, 2025. DOI: 10.1109/ACCESS.2025.3645115. [Online]. Available: https://doi.org/10.1109/ACCESS.2025.3645115.

[43] X. Wang, W. Liu, H. Guan, and Y. Wang, "Cognitive load recognition model based on improved attention mechanism and domain-adversarial network," in *International Conference on Information, Computing and Technology*, Springer, 2025, pp. 133–148.

[44] X. Ye and K. I.-K. Wang, *Domain-adversarial anatomical graph networks for cross-user human activity recognition*, 2025. arXiv: 2505.06301 [cs.LG]. [Online]. Available: https://arxiv.org/abs/2505.06301.

[45] C. Zhang, "Research on cross - domain data adaptation technology in transfer learning," *ITM Web of Conferences*, vol. 78, Sep. 2025. DOI: 10.1051/itmconf/20257804034.