

Toward Adaptive and User-Centered Intelligent Vehicles: AI Models
with Granular Classifications for Risk Detection, Cognitive
Workload, and User Preferences

by

Hyowon Lee

A thesis

presented to the University of Waterloo

in fulfillment of the

thesis requirement for the degree of

Doctor of Philosophy

in

System Design Engineering

Waterloo, Ontario, Canada, 2024

© Hyowon Lee 2024

Examining Committee Membership

The following served on the Examining Committee for this thesis proposal. The decision of the Examining Committee is by majority vote.

External Examiner	Alexander Crizzle Associate Professor
Internal-External Examiner	Benjamin Thompson Professor
Supervisor(s)	Siby Samuel Assistant Professor
Internal Member	Shi Cao Associate Professor
Internal Member	Kumaraswamy Ponnambalam Professor

Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

As artificial intelligence (AI) increasingly integrates into our transportation systems, intelligent vehicles have emerged as research topics. Many advancements aim to enhance both the safety and comfort of drivers and the reliability of intelligent vehicles. The main focus of my research is addressing and responding to the varying states and needs of drivers, which is essential for improving driver-vehicle interactions through user-centered design. To contribute to this evolving field, this thesis explores the use of physiological signals and eye-tracking data to decode user states, perceptions, and intentions. While existing studies mostly rely on binary classification models, these approaches are limited in capturing the full spectrum of user states and needs. Addressing this gap, my research focuses on developing AI-driven models with more granular classifications for cognitive workload, risk severity levels, and user preferences for self-driving behaviours. This thesis is structured into three core domains: collision risk detection, cognitive workload estimation, and perception of user preferences for self-driving behaviours. By integrating AI techniques with multi-modal physiological data, my studies develop ML (Machine Learning) models for the domains introduced above and achieve high performance of the ML models. Feature analytical techniques are employed to enhance model interpretability for a better understanding of features and to improve the model performance. These findings pave the way for a new paradigm of intelligent vehicles that are not only more adaptive but also more aligned with user needs and preferences. This research lays the groundwork for the future development of user-centered intelligent companion systems in vehicles, where adaptive, perceptive, and interactive vehicles can better meet the complex demands of their users.

Acknowledgments

First and foremost, I offer my deepest gratitude and glory to God, who has always watched over me and guided me safely through every step of this journey. Through the most challenging moments, His words have been my unwavering source of strength. Without them, I would not have been able to complete my PhD.

"I am with you and will watch over you wherever you go, and I will bring you back to this land. I will not leave you until I have done what I have promised you." – *Genesis 28:15*

"Sorrowful, yet always rejoicing; poor, yet making many rich; having nothing, and yet possessing everything." – *2 Corinthians 6:10*

I would also like to express my heartfelt appreciation to everyone who has supported and encouraged me along the way. While I wish I could name each and every one of you, I am especially grateful to: my father, Cheonho Lee, my mother, Jungrye Kim, and my sister, Yuri Lee, whose unwavering love and support have been my greatest strength. My supervisors who have guided me along the way from undergraduate to PhD graduations, particularly my supervisors, Prof. Ocktaeck Lim, Prof. Jaal Ghandhi, and Prof. Siby Samuel, for their invaluable advice and encouragement. Dr. Byounghoon Kim, who provided me with guidance and support during a time when I was considering changing my field of study. My aunts, Kim Junghee and Kim Youngran, who have always believed in me and cheered me on. My dear friends, Hosung Bang, Huiwon Park, and Hyunbo Shim, whose encouragement has been a source of motivation and strength. Finally, I thank Wendy Shin, my beloved girlfriend, for always being by my side as a supporter.

From the moment I set foot in graduate school, life seemed to unfold as a series of trials, each one testing my resolve. There were moments when I felt as though the world was whispering to me, urging me to let go, to surrender to the weight of disappointment. No matter how much effort I poured in, my plans did not always unfold as I had hoped. PhD admissions in a new field felt like an unrelenting climb, each rejection another weight upon my shoulders, while financial worries cast a lingering shadow over my days. But nothing shook me quite like failing my qualifying exam, it rattled me to my very core, leaving me adrift in doubt, questioning everything: my choices, my future, and whether I had the strength to keep moving forward. And yet, the challenges did not stop there. They came in waves, each one pushing me deeper into uncertainty. I ventured into a new field, only to be met with rejection

after rejection, each one carving another scar into my sense of purpose. I found myself lost without the guidance I had hoped for, and the absence of a supervisor and indifference left me adrift in an ocean of doubt. Research topics shifted, ideas crumbled, and at one point, I even had to change advisors entirely, another detour in a journey already filled with unexpected turns. But among all these hardships, the battle with panic disorder was the most merciless. It did not simply challenge me; it consumed me. There were days when even breathing felt like an impossible task, when the walls around me seemed to close in, and when the thought of moving forward felt unbearable. More than once, I stood at the edge of giving up, exhausted by the weight of it all. Yet, somewhere in the depths of those struggles, I found something I never expected, resilience. I learned that strength is not the absence of hardship but the willingness to rise despite it. That perseverance is not about avoiding failure but about embracing it, learning from it, and continuing forward. That even when the road feels endless, even when the falls are too many to count, I can still stand, dust myself off, and take another step. Now, as I place the final period on this chapter of my life, I do so not with exhaustion, but with quiet confidence. The journey has been long, but I step forward with a heart full of lessons and a mind open to new possibilities. There are more challenges ahead, no doubt—but this time, I walk toward them not with fear, but with the unwavering belief that I will keep going, no matter what lies ahead.

Table of Contents

Examining Committee Membership.....	ii
Author's Declaration	iii
Abstract	iv
Acknowledgments	v
List of Figures	xiv
List of Tables.....	xix
List of Abbreviations.....	xxi
Chapter 1 Introduction.....	1
Chapter 2 Background.....	3
2.1 Overview	3
2.2 Error and Risk Perception and Responses from Physiological Signals.....	4
2.2.1 Cognitive Processing of Error and Risk Perception	5
2.2.2 Error Type and Risk Severity Level	6
2.2.3 Requirements of Physiological Recording Systems for Perceiving Error/Risk.....	7
2.3 Cognitive Workload and Mental Fatigue	8
2.3.1 Associations Between Mental Workload and Mental Fatigue.....	9
2.3.2 Characteristics of Mental Workload and Mental Fatigue in Physiological Signals	10
2.3.3 Requirements of Physiological Recording Systems for Detecting Mental Workload and Mental Fatigue.....	11
2.4 User Preference for Self-Driving Behaviours	12
2.4.1 Individual Perception.....	12
2.4.2 Varied Driver Preferences with Multiple Factors.....	12
2.4.3 Cognitive Processing of User Preference	13
2.4.4 Key Associations Between User Preferences and Cognitive Process	14

2.4.5 Requirements of Physiological Recording Systems in Perceiving User Preference in Self-Driving Behaviours	15
2.5 Research Questions	16
2.5 Toward a Meta-Framework by Integrating User Perception Models	18
Chapter 3 Methodology.....	20
3.1 Overview of Chapters.....	20
3.2 Recording Systems for Physiological Signals	20
3.3 List of ML Models.....	22
3.4 Features Extracted from Physiological Signals	23
3.4.1 EEG Channels and Various Frequency Bands in EEG Signals.....	23
3.4.2 Pupil Data from Eye-Tracking System.....	25
3.4.3 Other Physiological Signals: PPG, EDA and Temperature	25
3.4.4 Extracted Features	26
Chapter 4 Artifact Removal of Physiological and Eye-Tracking Signals	28
4.1 Artifact Removal of EEG Signals	28
4.1.1 Two Common Algorithms for Filtering Artifacts in EEG Signals.....	29
4.1.1.1 Common Average Reference (CAR).....	29
4.1.1.2 Revised Common Average Reference (R-CAR).....	30
4.1.1.3 Independent Component Analysis (ICA)	32
4.1.1.4 Automatic Independent Component Analysis (A-ICA)	33
4.1.2 Comparison of Preprocessed Signals	35
4.2 Artifact Removal of Eye-Tracking Data	39
4.2.1 Hampel Filter.....	39
4.2.2 Preprocessing Procedure for Eye-Tracking Signals	40
4.3 Artifact Removal of Other Physiological Signals	43

4.3.1 Blood Volumn Pulse (BVP)	44
4.3.2 Electrodermal Activity (EDA).....	46
4.3.3 Temperature.....	48
4.4 Computational Complexity and Scalability.....	48
4.5 Summary	49
4.6 Limitation	50
Chapter 5 Detection of Braking Activation and Different Deceleration Degrees Related to Vehicle Speed at Collision.....	51
5.1 Experiment Design	51
5.1.1 Braking Time and Distance	52
5.1.2 Design of Traffic Scenarios.....	54
5.1.3 Participants	59
5.1.4 Experiment Procedure	60
5.2 Data Analysis.....	60
5.2.1 Signal Preprocessing and Feature Extraction	60
5.2.2 Feature Epoch.....	60
5.2.3 Class Imbalance.....	62
5.2.4 Feature Analysis Methods	62
5.2.5 Various Granularity Levels in Three Traffic Environments.....	66
5.3 Building User Perception Models for Braking Activation/Deceleration and Other Non-Braking Actions.....	68
5.3.1 Performance of the User Perception Models (Traffic Environment 1)	69
5.3.2 Performance of the User Perception Models (Traffic Environment 2)	71
5.3.3 Performance of the User Perception Models (Traffic Environment 3)	72
5.4 Feature Importance Analysis.....	74

5.4.1 Feature Analysis in Scenarios: Collision with Pedestrian(s)	75
5.4.2 Feature Analysis in Scenarios: Collision with Vehicle(s)	82
5.5 Discussion	89
5.5.1 Potential Applications of Subjectively Perceived Risks by Physiological Sensing Methods	90
5.5.2 Optimizing Temporal Resolution in Critical Situation Detection: Comparing EEG, Eye- tracking, and Other Physiological Signals.....	95
5.6 Limitations.....	96
5.7 Conclusion.....	98
Chapter 6 Workload Estimation Models Using Physiological and Eye-Tracking Signals.....	99
6.1 Experiment Design	99
6.1.1 NASA Task Load Index (NASA TLX).....	99
6.1.2 Data Collection.....	100
6.1.3 Participants	100
6.2 Data Analysis.....	101
6.2.1 Feature Epoch.....	102
6.2.2 Feature Type.....	102
6.3 Machine Learning Approaches to Building Workload Prediction Model.....	102
6.3.1 Supervised Learning.....	103
6.3.1.1 Supervised Learning Algorithm: Regression	100
6.3.1.2 Performance Evaluation of Regression Models	106
6.3.2 Unsupervised Learning.....	107
6.3.2.1 Unsupervised Learning Algorithm: Clustering	108
6.3.2.2 Performance Evaluation of Clustering Algorithms	109
6.3.3 Semi-Supervised Learning	111

6.3.3.1	Semi-Supervised Learning Methods	111
6.3.3.2	Perofmrnce Evaluation of Semi-Supervised Learning Methods	111
6.4	Resutls: NASA TLX and Performance of Machine Learning Approaches.....	112
6.4.1	NASA TLS Result	112
6.4.2	Aggregation of NASA-TLX Results from Two Different Experiments	116
6.4.3	Supervised Learning Result.....	117
6.4.4	Unsupervised Learning Result.....	119
6.4.4.1	Optimal Cluster Number	122
6.4.4.2	Discussion	122
6.4.5	Semi-Supervised Learning Result	123
6.4.5.1	Condition 1: Binary Classification of Workloads (2 Clusters).....	123
6.4.5.2	Condition 2: Multi-class Classification of Workloads (4 Clusters).....	124
6.5	Feature Analysis of Workload Estimation Models.....	128
6.5.1	Binary Workload Classification	128
6.5.2	Multi-class Workload Classification (4 Classes ~ 4 Clusters)	128
6.6	Adaptive Models of Workload Estimation.....	129
6.7	Discussion	130
6.8	Limitations.....	131
6.9	Conclusion.....	132
Chapter 7	Perception Model of User Preference for Self-Driving Behaviours	134
7.1	Study Design	134
7.1.1	Selected Traffic Situations.....	134
7.1.2	Preference Options for Self-Driving Behaviours	139
7.1.3	Participants	140
7.1.3.1	Survey Participants.....	140

7.1.3.2 Experiment Participants.....	140
7.1.4 Data Collection Procedure.....	141
7.1.4.1 Traffic Situation Recording System	141
7.1.4.2 Survey Procedure.....	142
7.1.4.3 Experiment Procedure	143
7.2 Data Analysis.....	143
7.2.1 Survey Analysis.....	143
7.2.2 Experiment Data Analysis	145
7.3 Survey Result.....	148
7.3.1 Traffic Complexity Effect on Difficulty Levels	149
7.3.2 Traffic Complexity Effect on Preference Modes.....	150
7.3.3 Statistical Tests: Binary Preferences and Difficulty Levels	152
7.3.4 Preference Modes and Preferred Actions	153
7.3.5 Correlation Analysis for Preference and Individual Perception/Experience	156
7.3.6 Interaction Effects: Preference and Individual Perception/Experience	157
7.3.7 Interaction Effects: Features Extracted from Survey Responses	159
7.3.8 Discussion.....	165
7.4 Predictive Model of User Preference Using Survey Data	166
7.4.1 Model Selection and Optimization Process	166
7.4.2 Preference Classification: Granularity of Driver Preferences	168
7.4.3 Classification Results of Predictive Models.....	169
7.4.4 Feature Importance Analysis	170
7.4.5 Discussion.....	174
7.5 Predictive Model of User Preference Using Experiment Data	175
7.5.1 Performance of User Preference Models.....	175

7.5.2 Feature Importance Analysis	177
7.5.3 Discussion.....	180
7.6 Conclusion	181
Chapter 8 Conclusion	182
8.1 Contributions	182
8.2 Conceptual Integration of User Perception Models.....	184
8.3 Future Work	186
8.3.1 Suggestions for Furthre Studies.....	186
8.3.2 Long-Term Vision	191
References	192
Appendix A Lazy Predict: 30 ML Models	215
Appendix B Extracted Features	217

List of Figures

Figure 1	Integrated user perception model.....	4
Figure 2	Selected EEG channels in the existing studies.....	24
Figure 3	Selected channels for re-referencing of Cz channel (Left) and various sets of selected channels that each region represents a subset of channels for re-referencing of a channel (Right).....	30
Figure 4	A schematic diagram of preprocessing with the R-CAR algorithm	31
Figure 5	A schematic diagram of EEG artifact removal with blind source separation in Independent Component Analysis [184].....	32
Figure 6	Variance ratios of independent components	34
Figure 7	A schematic diagram of preprocessing with the A-ICA algorithm.....	35
Figure 8	EEG amplitudes from raw signals, CAR, R-CAR, ICA, and A-ICA	36
Figure 9	Filtered EEG amplitude's variances from raw, CAR, R-CAR, and A-ICA algorithms with a bandpass filter over 300 trials: Left (variances in fixations) and Right boxplot (variances in saccades).....	38
Figure 10	A schematic diagram of preprocessing eye-tracking data.....	42
Figure 11	Raw pupil area data (Left) and processed interpolation of pupil area data in saccades (Right).....	42
Figure 12	Hampel identifier: outlier removal.....	43
Figure 13	Empatica E4 specification [192]	44
Figure 14	Detection theory of PPG sensor and its output [193].....	42
Figure 15	PPG waveform with annotated key points: Diastolic point (local minima), Systolic point (local maxima), Dicrotic notch, and Dicrotic wave. These points provide valuable information for assessing cardiovascular function, including the estimation of heart rate, vascular resistance, and potential cardiac abnormalities (Original image from [196]) ...	46

Figure 16	PPG waveform with regular and irregular diastolic points.....	46
Figure 17	EDA signals: phasic activations are circled (Top) and slow changes with no phasic significant activity (Bottom) [196]	47
Figure 18	CARLA simulation environment.....	52
Figure 19	Braking distance and time at various vehicle speeds	54
Figure 20	Captured scenes from experiments	56
Figure 21	Traffic scenario design/procedure.....	61
Figure 22	Illustration and process of the oversampling technique (ADASYN).....	63
Figure 23	Performance across various granularity levels in Traffic Environment 1.....	69
Figure 24	Performance of pedestrian perception models	70
Figure 25	Performance of pedestrian/vehicle speed perception models	70
Figure 26	Performance across various granularity levels in Traffic Environment 2.....	71
Figure 27	Performance across various granularity levels in Traffic Environment 3.....	72
Figure 28	Performance of pedestrian perception models	73
Figure 29	Performance of pedestrian/vehicle speed perception models	74
Figure 30	Feature analysis in Traffic environment 1: pedestrian types (Left) and # of pedestrians (Right).....	76
Figure 31	Feature analysis of three cases in Traffic environment 1: Alerting, Error and No error .	77
Figure 32	Feature analysis of three factors in Traffic environment 1: No Collision, Collision and Collision severity	78
Figure 33	Feature analysis of two factors in Traffic environment 3: Pedestrian types (Left) and Number of pedestrians (Right).....	79
Figure 34	Feature analysis of three cases in Traffic environment 3: Alerting, Collision and No collision.....	80
Figure 35	Feature analysis of three factors in Traffic environment 3: No collision, Collision and Collision severity	81

Figure 36	Feature analysis of three factors in Traffic environment 1: Alerting, No collision and Collision.....	83
Figure 37	Feature analysis of three factors in Traffic environment 1: No collision, Collision and Shoulder	84
Figure 38	Feature analysis of three factors in Traffic environment 2: Alerting, AV and Collision.	85
Figure 39	Feature analysis of three factors in Traffic environment 2: AV, Collision and SV2.....	86
Figure 40	Feature analysis of three factors in Traffic environment 2: No collision, Alerting, AV .	87
Figure 41	Feature analysis of three factors in Traffic environment 2: No collision, Collision and Collision severity	88
Figure 42	EEG systems with improved convenience: MindWave from NeuroSky, BR8+ from BRI and EPOC from Emotiv [246]	98
Figure 43	A schematic diagram of data acquisition for participant fatigue states	100
Figure 44	Epoch extraction based on prediction duration.....	102
Figure 45	Six dimensions of NASA-TLX in pre-, intermediate- and post-experiment	113
Figure 46	Difference of scale responses in each dimension between intermediate- and post-experiment NASA-TLX.....	114
Figure 47	Total workload and mental workload for all participants	115
Figure 48	Comparison of experiment completion times in two experiments.....	116
Figure 49	Performance of the selected regression models (MSE, MAE and R^2).....	118
Figure 50	Prediction analysis for Bagging Trees for original features and Gaussian Process Regression for reduced features by PCA	119
Figure 51	Comparison of clustering performance results in various cluster numbers (Original features vs Reduced features by PCA).....	121
Figure 52	Clustering examples based on the number of clusters (2 and 4 clusters).....	123
Figure 53	Decision process of binary workload classes.....	124
Figure 54	Classification performance based on matched workload classes with clusters	125

Figure 55	Decision process of multi-workload classes (Four workload levels).....	126
Figure 56	Illustration of Clusters: Mapping Clusters to Workload Classes for Four workload levels	126
Figure 57	Classification performance based on matched workload classes with clusters	127
Figure 58	Percentage of binary workload levels (2 Levels) and multiple workload levels (4 levels)	127
Figure 59	Feature importance results in binary and multiple classifications: Two workload classes (Left) and Four workload classes (Right)	130
Figure 60	Areas where selected videos were recorded in the city of Toronto (Left) and the regional municipality of Waterloo (Right).....	134
Figure 61	The comparison of the average traffic counts per hour in each situation of the two regions (S#: Situation number)	138
Figure 62	AKASO V50 camera mounted on dashboard	141
Figure 63	Video session procedure	142
Figure 64	Two epoch durations for first impression and main traffic events (Frames are extracted from ‘S1: Police in high traffic volume, Toronto’).....	146
Figure 65	Age differences in numbers of advanced features experienced (demographic info.), number of “Prefer” choices, and average difficulty levels in 18 situations	149
Figure 66	Proportion differences between the two traffic complexities (% difference = % Low traffic complexity – % High traffic complexity)	151
Figure 67	Preferences for ‘being informed’ action and human-engaged actions in the two age groups (Human-engaged actions included ‘Inform-Consent’ and ‘Alert-Handover’)...	154
Figure 68	Proportion differences of being informed between the age groups in two traffic complexities (difference % = Y-M group – Old group)	154
Figure 69	Proportion differences of human-engaged actions between the age groups (% Y-M – Old) in two traffic complexities	155
Figure 70	Correspondence analysis of two groups’ preferred actions for Toronto/Waterloo and roundabout situations	156

Figure 71	Decision tree analysis for the Y-M group (Dependent variable: Binary preference)	160
Figure 72	Decision tree analysis for the Older group (Dependent variable: Binary preference) ...	161
Figure 73	Decision tree analysis for the Y-M group (Dependent variable: Multiple actions).....	162
Figure 74	Decision tree analysis for the Older group (Dependent variable: Multiple actions).....	164
Figure 75	Model training procedure.....	167
Figure 76	Shapley values of the top eight features in ML models with all granularity levels	174
Figure 77	Classification performance of user preference models. (Left) 2 options: ‘Prefer or not prefer self-driving’ and (Right) 5 options: Act-Inform, Inform-Consent, etc.....	176
Figure 78	Classification performance of user preference models (Granularity level 2~5)	177
Figure 79	Comparison of feature analysis from granularity level 2 to 5.....	179
Figure 80	A schematic diagram of the integrated user perception model	185
Figure 81	How ADAS works [287]	186
Figure 82	New integrated user perception model with traffic information	188
Figure 83	Post-ride AV tuner to update user preferences on the decision-making model of AV behaviours.....	189
Figure 84	Advanced traffic environment perception through Vehicle-to-Everything communications	190

List of Tables

Table 1	Contradictory results of power in EEG frequency bands.....	10
Table 2	Enobio 32 technical specifications.....	21
Table 3	Ergoneers Dikablis 3 Eye-tracker technical specifications.....	21
Table 4	Empatica E4 technical specifications.....	22
Table 5	EEG frequencies used across this thesis	25
Table 6	Eye-tracking data used across this thesis	25
Table 7	List of features extracted from physiological signals	26
Table 8	Sampling rates of physiological signals used across this thesis.....	28
Table 9	Average number of selected ocular components by ICA and A-ICA in 20 trials.....	34
Table 10	Average and standard deviation of filtered EEG amplitude's variances from four different preprocessing conditions in fixations and saccades: Raw, CAR, R-CAR, and A-ICA.....	37
Table 11	Values for the box plots from filtered EEG amplitude's variances from four different preprocessing conditions in fixations and saccades: Raw, CAR, R-CAR, and A-ICA ...	38
Table 12	Three traffic environments with various factors	55
Table 13.1	Defined classes from pedestrian-related situations in Traffic environment 1.....	58
Table 13.2	Defined classes from pedestrian-related situations in Traffic environment 2.....	58
Table 13.3	Defined classes from pedestrian-related situations in Traffic environment 3.....	59
Table 14.1	Classes in various granularity levels (Traffic Environment 1)	66
Table 14.2	Classes in various granularity levels (Traffic Environment 2)	67
Table 14.3	Classes in various granularity levels (Traffic Environment 3)	67
Table 15	Calculated weights for six dimensions in NASA-TLX.....	115
Table 16	Counterbalanced situations and their descriptions.....	135

Table 17	AV behaviour options designed by Park et al. [260] and Billings [261]	139
Table 18	Categorization of preferred actions based on different aspects: user preference, informing action, and action agent [264]	139
Table 19	Age and gender information	140
Table 20	Details of one-time and situation questionnaires	142
Table 21	Statistical tests were used in this study	144
Table 22	Baseline of statistical power analysis (Power = 0.8, Effective size $w = 0.25$ or 0.3)	144
Table 23	Features extracted from the survey responses and the traffic situation information.....	144
Table 24	Average data for survey responses.....	149
Table 25	Results of the Chi-square test: Traffic complexity and difficulty levels.....	150
Table 26	Chi-squared test: Traffic complexity and preference mode	151
Table 27	Chi-squared test: mode preference and difficulty levels.....	152
Table 28	Chi-squared test: preferred actions of high/low traffic conditions in each group	153
Table 29	Spearman's rank correlation coefficient	156
Table 30	Significance results of interaction effects between the independent variables and the dependent variable shown in Table 21 ($N = 85$ for each variable)	158
Table 31	Optimized hyperparameter ranges for user preference ML models at the four granularity levels (from the granularity level 2 to 5).....	167
Table 32	Preferred actions in various granularity levels [286]	169
Table 33	The comparison between the baseline model and the final model performances from the granularity level 2 to 5 (Optimized model results)	170
Table 34	Descriptive statistics: Shapley values of the top eight features in	173
Table 35	Feature importance ranks of the three age groups with four granularity levels from Level 2 to Level 5 evaluated by Shapley values	173
Table 36	List of surrounding and upcoming traffic environment	187
Table 37	Detection of fatigued states of drivers	188

List of Abbreviations

AI	Artificial Intelligence
AASHTO	American Association of State Highway and Transportation Officials
ABS	Anti-lock Braking System
ACC	Anterior Cingulate Cortex
ADAS	Advanced Driver Assistance System
A-ICA	Automatic Independent Component Analysis
ANOVA	Analysis of Variance
ANS	Autonomic Nervous System
BCI	Brain-Computer Interface
BSS	Blind Source Separation
BVP	Blood Volume Pulse
CAR	Common Average Reference
CHI	Calinski-Harabasz index
CoSI	Cognitive Style Indicator
CV	Cross-Validation
DBI	Davis-Bouldin Index
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
EDA	Electrodermal activity
EEG	Electroencephalogram
ECG	Electrocardiogram
ERN	Error-Related Negativity
ErrPs	Error-related Potentials
ERPs	Event-Related Potentials
EOG	Electrooculographic
ESC	Electronic Stability Control
GMM	Gaussian Mixture Model
GPR	Gaussian Process Regression
GSR	Galvanic Skin Response

GVWR	Gross Vehicle Weight Rating
HRV	Heart Rate Variability
IBI	Inter-Beat Interval
ICA	Independent Component Analysis
IQR	Interquartile Range
LASSO	Least Absolute Shrinkage and Selection Operator
LGBM	Light Gradient-Boosting Machine
LIF	Lower Inner Fence
MAD	Median Absolute Deviation
MAE	Mean Absolute Error
MCQ	Multi-Choice Question
ML	Machine Learning
MLP	Multi-Layer Perception
MSE	Mean Squared Error
NASA-TLX	National Aeronautics and Space Administration-Task Load Index
Ne	Error-related Negativity
NHTSA	National Highway Traffic Safety Administration
OLS	Ordinary Least Squares
PCA	Principal Component Analysis
PCD	Pupil Center shift Degree
PCM	Pupil Center shift Magnitude
Pe	Error-related Positivity
PPG	Photoplethysmogram
PS	Pupil Size
Q1~3	Quartile 1~3
QoE	Quality of Experience
R-CAR	Revised Common Average Reference
R ²	Coefficient of Determination
ROC	Receiver Operating Characteristic
SD	Standard Deviation
SGD	Stochastic Gradient Descent

SHAP	Shapley Additive exPlanations
SVC	Support Vector Classifier
SVM	Support Vector Machine
UIF	Upper Inner Fence
WCSS	Within-Cluster Sum of Squares
XGBoost	eXtreme Gradient-Boosting
Y-M	Young-Middle aged group (23-44 years old)

Chapter 1. Introduction

Rapid advancements in artificial intelligence (AI) and breakthroughs in physiological signal processing are revolutionizing advanced vehicle design. Central to this transformation is the development of user-centred systems for advanced vehicles such as semi-autonomous vehicles (semi-AVs). Unlike traditional automotive systems that rely solely on mechanical and electronic components, future vehicles will be equipped with intelligent systems capable of understanding, predicting and responding to human physiological and cognitive states. This evolution is more than just a technological upgrade; it represents a paradigm shift towards creating vehicles that act as intelligent partners rather than mere tools for transportation. The aim is to foster a more interactive relationship between humans and machines. For instance, vehicles that can effectively communicate with users by understanding their emotional and cognitive states and making decisions aligned with user preferences will play a crucial role in future mobility solutions. Achieving this requires a comprehensive framework that leverages physiological signals. Integrating physiological data such as eye-tracking, PPG (photoplethysmography), EDA (electrodermal activity) and EEG (electroencephalogram) with AI-driven models can enable sophisticated user perception systems that adapt in real-time to individual user needs. The motivation for this shift stems from demands for vehicles that provide safer and more efficient transportation and offer highly personalized experiences. Developing AI-driven user perception models requires a strong foundation in preprocessing and artifact removal of physiological signals, as these steps are critical for extracting reliable data features essential to model development. Chapter 4 emphasizes the integration of diverse physiological signals, such as EEG and eye-tracking, to apply advanced artifact removal methods that improve feature extraction. This chapter demonstrates that integrating multiple modalities refines data quality and improves the quality of features extracted from physiological signals. These features form the cornerstone for developing user perception models such as error and risk detection and user preferences, emphasizing multi-modal data integration's significance for advanced vehicle systems.

Building on this foundation, Chapter 5 explores the development of AI-driven models for detecting collision risks and assessing their severity in different traffic environments. The study employs a range of machine learning (ML) techniques to achieve high classification performance, with precision, recall, F1 scores, and accuracy consistently exceeding 0.9 in binary classifications of collision scenarios. These models demonstrate robust performance even as classification complexity increases, highlighting their adaptability. However, the research emphasizes the importance of context, showing that model stability and effectiveness can vary depending on specific traffic conditions, such as pedestrian versus vehicle collisions. This underscores the need for context-aware AI models that can dynamically adjust to different traffic environments for accurate risk assessments.

Chapter 6 explores user preferences for self-driving behaviours, with a focus on understanding and addressing the need for personalized and adaptive autonomous vehicle systems. This chapter emphasizes how semi-AVs can be tailored to accommodate individual user needs and preferences by integrating personalized ML models that adjust automation levels based on traffic conditions and user-defined parameters. This adaptability enhances user satisfaction and trust, promoting a more user-centred approach to autonomous vehicle design. The findings suggest a paradigm shift towards advanced user-centred AV systems that deliver accurate, reliable and customized driving experiences in increasingly complex environments.

Continuing the theme of user-centred design, Chapter 7 investigates the estimation of cognitive workload using physiological signals and eye-tracking data. By employing supervised, unsupervised and semi-supervised learning methods, the study develops robust models for predicting mental workload levels. The findings emphasize the importance of effective feature selection and clustering techniques in enhancing model performance. This comprehensive approach contributes to workload assessment in autonomous systems by providing a foundation for more nuanced and responsive vehicle interactions that anticipate and adapt to users' cognitive states.

This thesis represents a significant step toward integrating physiological signal analysis and AI-driven user perception models into advanced vehicle design. By combining signal processing, risk detection, user preference modelling and workload estimation, this work lays the groundwork for a new era of intelligent and adaptive vehicle systems. The goal is to create semi-AVs that are highly functional and deeply attuned to human needs, providing safer, more efficient and personalized driving experiences. As the field evolves, the insights and models developed here will be a critical foundation for future advancements in user-centred autonomous vehicle design.

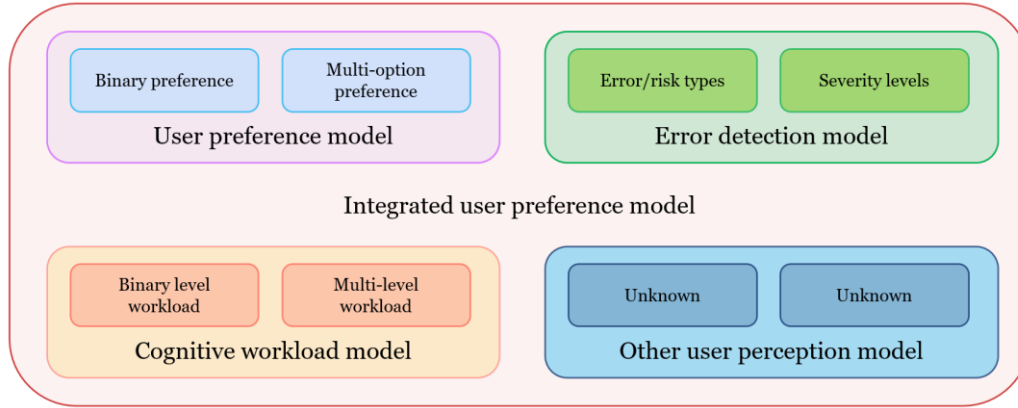
Chapter 2. Background

2.1 Overview

This chapter provides the background necessary to understand the significant components of each research topic in this thesis, indicating our research at the intersection of human-machine interaction, physiological signals, user-centred design and AI-driven models for advanced vehicles. We first explore human responses to error detection as a form of physiological signals focusing on cognitive processing of error/potential risks (Section 2.2: Error/risk Perception and Responses from Physiological Signals). This section also discusses the specific requirements for physiological recording systems to capture features extracted from physiological signals when a human responds to a stimulus (Section 2.2.3: Requirements of Physiological Recording Systems for Perceiving Errors). In the next section, we provide the background regarding cognitive workload and mental fatigue that have been explored in driver-vehicle interaction. These topics are among the central states in drivers that can lead to potentially risky situations due to impairments in driving performance. We first investigate distinct characteristics of cognitive workload reflected in physiological signals such as EEG, eye-tracking and heart rate features (Section 2.3.1: Characteristics of Cognitive Workload in Physiological Signals). Then, similar to Section 2.2.3, we identify the necessary technical specifications for such physiological recording systems for capturing features that can be built in AI-driven models (Section 2.3.2: Requirements of Physiological Recording Systems for Detecting Cognitive Workload). For the next section, we investigate user preference for self-driving behaviours which is another major focus of this thesis. We first review individual perceptions of self-driving technology, e.g., how individual perceptions and preferences are differentiated by environmental factors (Section 2.4.1: Individual Perceptions on Self-Driving). This analysis covers how preferences vary according to factors like age, experience and traffic conditions. Then, the following section provides more scientific references regarding various factors that can shape driver preferences (Section 2.4.2: Varied Driver Preferences with Multiple Factors). We further explore the associations between user preferences and cognitive processes, providing insights into how these factors interact to shape user preferences (Section 2.4.4: Key Associations Between User Preferences and Cognitive Processes). Then, as earlier we did in other topics above, we specify the technical requirements for systems that perceive user preferences through physiological signals in self-driving contexts (Section 2.4.5: Requirements of Physiological Recording Systems for Perceiving User Preferences in Self-Driving Behaviours). Figure 1 presents an integrated model combining three user perception models (Error/Potential risk detection, cognitive workload estimation and user preference prediction) into a unified framework. By synthesizing insights from these research areas, this chapter

establishes a foundational framework for developing AI-driven user perception systems for advanced vehicles, suggesting integration of the models for error detection, cognitive workload and user preference.

Figure 1. Integrated user perception model



2.2 Error and Risk Perception and Responses from Physiological Signals

An error in the context of driving is defined as a deviation from intended actions or decisions that lead to undesirable outcomes, such as misjudgments or lapses in attention, and is a major contributor to traffic incidents [1, 2]. On the other hand, risk refers to the potential for exposure to harm or loss, often quantified by the likelihood and severity of adverse events. Risk perception is highly subjective, influenced by individual differences, such as prior experiences and personal biases, which affect how drivers assess and respond to hazards [3, 4, 5]. Errors and risks are closely related, as driver errors can elevate risks by increasing the probability of hazardous situations, while high-risk scenarios can heighten the likelihood of errors due to increased cognitive demands. Studies reveal a bidirectional relationship, showing that perception errors directly impact evaluations of risk levels [3]. For example, in driving situations, collision risks are closely related to whether braking actions are adequately activated. When the braking system fails to activate correctly (referred to as a braking system error), driver perception can become wavering as ambiguous and complex states between perceiving an error and a risk. This ambiguity arises because inappropriate or absent braking activations are caused by a malfunctioning braking system in autonomous driving conditions. Before recognizing a situation as having a high collision risk, users or drivers might initially perceive these issues as system errors. These system errors are often related to potential collision risks that can lead users to transition naturally between cognitive processing of error and risk. For example, as the distance to a potential collision vehicle shortens, cognitive perception may shift from focusing on system errors to prioritizing the perception of collision risk. In other words, if the brakes do not activate within a safe distance, drivers may transition their perception from error to risk. However, research is still

limited to when precisely this perceptual transition occurs. Thus, we conceptualize that errors and risks can have a closely intertwined cognitive processing. In autonomous driving conditions, delayed activation of the braking system due to late braking or deceleration timing may cause users or drivers to prioritize the perception of an error in their cognitive processing.

2.2.1 Cognitive Processing of Error and Risk Perception

Human cognition studies are an interdisciplinary field where various disciplines intersect, such as psychological and physiological perspectives, human behaviours and neurosciences. Cognition encompasses diverse mental processes, e.g., attentive states, perception, etc. In human-computer interaction, these processes have been estimated by using various features extracted from physiological signals. In scenarios involving error and risk perception, some features extracted from physiological signals may serve as markers. Humans can cognitively perceive both errors and risks and this cognition can stimulate their physiological systems. These are mediated by the autonomic nervous system (ANS). This system involves both the sympathetic and parasympathetic systems. For example, the ANS responds to stress, a human state, and its responses are reflected in physiological signals [6, 7, 8]. Therefore, the ANS plays a pivotal role in modulating the body's responses to perceived errors [8, 9]. In the ANS, the sympathetic system may induce a 'fight or flight' response which is characterized by heightened alertness. It can be highly correlated with the state that facilitates rapid perception of and responses to errors and risks. In contrast, the parasympathetic system promotes a 'rest and digest' state associated with more deliberate and reflective responses to errors [10]. The modulation between these sympathetic and parasympathetic systems can be shown in various physiological signals, e.g., electroencephalogram (EEG), galvanic skin response (GSR), electrocardiogram (ECG), photoplethysmogram (PPG) and pupil data from an eye-tracking system [8, 11, 12, 13, 14, 15, 16]. Regarding EEG, the anterior cingulate cortex (ACC) is associated with its activation during error detection [17, 18, 19]. This activation is often reflected in EEG signals as error-related negativity (Ne) and positivity (Pe) potentials [21]. Additionally, the insular cortex which involves cognitive awareness, shows increased activity during autonomic responses. This indicates changes in human affective states during cognitive processing, e.g., errors [15]. As indicators for cardiovascular activity, ECG and PPG have been used to study ANS roles during cognitive processing. Both signal modalities can indicate subtle shifts in autonomic arousal associated with error detection when an individual perceives an error within a visual stimulus [22, 23]. These studies also point out that delayed and slower heart rates are evoked responses by humans to cognition of error. It underlies physiological changes in human arousal states. Another feature of heart rate, e.g., heart rate variability (HRV), is shown as distinctive changes evoked from the processing of unexpected stimuli and activation of cognitive control systems [8, 25]. Besides the above

indicators, GSR is also known as electrodermal activity (EDA) measures the electrical conductance of the skin [26]. Skin conductance is regulated by the sympathetic nervous system [24, 27]. It can serve as a direct marker of autonomic arousal in response to various stimuli, including perceived errors or visual anomalies [8, 20, 24, 27, 28]. During error cognition, results from Cheyne et al. (2000) indicate that observed changes in EEG signals are correlated with galvanic skin response [20]. Furthermore, as an indicator for cognition of a stimulus from eye-tracking data, pupil dilation has extensively been explored for the following stimulus responses: errors [8, 14, 17, 29], surprise [20] and noise intensity [21]. These stimulus responses point out a broader temporal modulation by ANS for cognitive and emotional states. Pupil dilation seems to be more significant indicator for errors compared to non-error stimuli [8, 17, 29]. At the single-trial level, Murphy et al. (2016) found a positive correlation between pupil dilation and erroneous stimulation [13]. In summary, physiological signals can be reliable indicators of responses to external stimuli and internal modulation for human cognition.

2.2.2 Error Type and Risk Severity Level

Errors can vary significantly in type and severity, ranging from minor lapses that have negligible consequences to critical mistakes that can lead to severe outcomes [22]. In the context of traffic accidents, errors can range from minor infractions, e.g., a slight deviation from a lane, to severe incidents, such as high-speed collisions leading to serious injury or fatalities [23]. Therefore, understanding of how humans perceive error types and severities can be a starting point for developing human perception systems and employing such systems into any form of human-computer interaction applications, e.g., various high-stakes environments such as traffic control, aviation and industrial operations. Especially, error severity perception in visual stimulation can be influenced by various factors, e.g., the salience of visual cues, their complexities and human cognitive loads [24]. For instance, drivers must continuously monitor their surroundings, identify potential hazards and assess the severity of various errors to make timely and appropriate decisions in traffic scenarios [25]. As we introduced earlier, physiological signals such as EEG, heart rate, and GSR can provide valuable insights into the cognitive states associated with error perception and response. Some studies point out that specific EEG patterns and elevated heart rate variability can indicate a higher perceived severity of errors [26, 27]. GSR measurements can reveal heightened arousal and stress responses when individuals encounter high-severity errors [28]. From eye-tracking data, we can enhance our understanding of how individuals perceive and respond to different error types beyond simply perception of error and non-error. Features from eye-tracking data can reveal patterns of visual attention such as fixation duration, saccades and gaze distribution that indicate how individuals process visual information and assess error severity [29]. For example, longer fixation durations on a particular area may

suggest increased cognitive processing and recognition of a high-severity error. One of my research projects highlighted the potential for multimodal approaches in error detection [30]. Furthermore, physiological signals can be used to monitor drivers' states and reactions to current traffic environments by detecting features correlated with potentially risky situations in the context of traffic accidents, e.g., slower reaction times and erroneous decisions before they lead to severe consequences. By understanding how drivers perceive and react to different error types, interventions can also be designed to mitigate risks and enhance safety. Accurate perception and classification of these errors can be used for effective error management and prevention strategies.

2.2.3 Requirements of Physiological Recording Systems in Perceiving Error/Risk

Understanding how humans perceive and respond to visual stimuli is crucial for developing effective error detection models using physiological signals. When the visual stimulus is a dynamic visual stimulus, the complexity of capturing precise human attentional shifts and error recognition moments increases significantly compared to static visual stimuli. A dynamic visual stimulus involves continuous frame changes and dynamic content that require careful attention to the timing of error perception and recognition. Video stimuli introduce more dynamic elements that can complicate identifying the timing when users recognize or perceive an error, unlike static conditions where there is only one target with no frame change [31, 32]. One of the primary challenges in using a dynamic visual stimulus for error detection is the lack of detailed information about human attentional focus, including the exact timing of when attention starts and ends on specific video segments. This absence of granularity in attentional data makes it difficult to build accurate error detection models. It can potentially degrade model performance and increase computational costs due to unnecessary feature extraction at non-optimal times [33]. To address this issue, precise estimation of epoch conditions such as the timing and duration of attentional focus is essential, particularly when recording physiological responses like EEG, GSR, PPG and ECG [34]. Eye-tracking systems are highly valuable in this context as they provide essential information on user gaze fixation timing and regions of interest during fixation duration. This information helps streamline computational processes by guiding the recording systems to focus on periods when users are overtly attending to specific video content [36, 37]. Studies by Ojsteršek and Topolšek [31] as well as Rai and Le Callet [32], highlight that eye-tracking can accurately capture overt visual attention by identifying where and how long users fixate on dynamic stimuli. Such insights are vital for synchronizing physiological data like EEG with moments of heightened attention, allowing for more precise extraction of relevant features from the recordings. This is supported by Gannouni et al. [33], who employed the zero-time windowing (ZTW) method to capture epochs in EEG data, demonstrating the importance of high temporal resolution in physiological recordings. Additionally,

physiological signals such as ECG, PPG and GSR provide valuable information but have certain limitations in the context of dynamic visual stimuli [39]. While HR-based features can reflect immediate responses to errors, they require longer time frames to capture meaningful changes, typically around 900 ms to 1500 ms [34]. In contrast, signals like EEG and pupil responses provide more immediate feedback as evoked responses begin around 200–300 ms after the onset of visual stimuli, making them more suitable for detecting rapid cognitive responses during dynamic video content [41, 42, 43, 44]. The complexity of using a dynamic visual stimulus lies in balancing the diverse temporal dynamics of different physiological signals to accurately classify error epochs in real-time. EEG and pupil responses that align more closely with the onset of visual changes offer a more direct and timely assessment of error perception than slower indicators like ECG, PPG and GSR. These findings suggest that leveraging signals that provide immediate physiological responses is essential for enhancing error detection models in dynamic visual environments [41, 42, 43, 44].

2.3 Cognitive Workload and Mental Workload

There are differences in definitions for cognitive workload and mental fatigue. Cognitive workload refers to the mental effort required to particularly process information and perform tasks, particularly those involving problem-solving and response to stimuli [45]. In contrast, mental workload is defined as "the degree of activation of a finite pool of resources, limited in capacity, while cognitively processing a primary task over time, mediated by external dynamic environmental and situational factors, as well as affected by definite internal characteristics of a human operator, for coping with static task demands, by devoted effort and attention" [46]. This definition highlights that mental workload not only includes cognitive demands but also integrates external environmental and situational factors, as well as internal characteristics of the individual. However, in some contexts, the terms, mental workload and cognitive workload, can be used interchangeably but encompasses a broader range of factors. These include not only task-specific cognitive demands but also emotional, physical, and environmental influences [47]. Both terms refer to the demand placed on an individual's cognitive resources when performing a task. They involve processes such as attention, memory, problem-solving, and decision-making [46, 48]. In the context of driving and driver state estimation, the terms "cognitive workload" and "mental workload" are often used interchangeably to describe the mental effort required to process driving-related tasks. For instance, a study on driver cognitive workload estimation refers to the process of monitoring a driver's cognitive status in real-time to assess their workload levels, using the term "cognitive workload" to encompass the mental demands placed on the driver [49]. Similarly, research examining drivers' mental workload during simulated driving scenarios employs the term "mental workload" to describe the cognitive demands experienced by drivers, indicating

that both terms are used to represent the mental effort associated with driving tasks [50]. These examples demonstrate that in driving research, "cognitive workload" and "mental workload" are often used interchangeably to describe the mental demands placed on drivers, reflecting the mental effort required to perform driving tasks effectively. Furthermore, various methods, including physiological (e.g., EEG, heart rate variability), behavioural (e.g., reaction times), and subjective (e.g., NASA-TLX) metrics, are used to assess both cognitive and mental workload [51, 52]. Self-assessment tools like NASA-TLX can not only measure overall workload by utilizing multiple items in the tool, but also measure mental demand as a separate item [53]. In addition, physiological recordings such as EEG, eye-tracking, GSR, ECG and PPG have enabled us to monitor user mental workload and fatigue states by correlating features from the recordings with the user states. EEG is one of the most commonly used methods for detecting cognitive workload and mental fatigue. Some studies have shown that specific EEG patterns, e.g., increased frontal theta and decreased occipital alpha power, are indicative of mental fatigue [54]. In the same context, eye-tracking can complement EEG by providing data on gaze patterns and pupil dynamics. Both EEG features and eye-tracking features can be used to monitor mental fatigue during the transition to mentally fatigued states [55]. Features related to heart activity can also be used to detect mental fatigue. For instance, HRV is associated with mental fatigue by indicating changes in autonomic nervous system activity [56]. GSR increases with stress and cognitive load as a direct measure of sympathetic nervous system activity [57]. Similarly, changes in body temperature and ECG parameters are also used to assess mental fatigue during visual display terminal tasks [58].

2.3.1 Associations Between Mental Workload and Mental Fatigue

Mental workload and mental fatigue are intrinsically linked, with sustained high mental workload frequently resulting in the onset of mental fatigue. Mental workload encompasses the cognitive and emotional resources required to perform a task and is shaped by factors such as task complexity, time constraints, and individual capabilities. Prolonged exposure to elevated mental workload can lead to mental fatigue and physical fatigue, which manifests as diminished alertness, impaired performance, and an overall sense of exhaustion [59, 60]. For example, studies have demonstrated a significant correlation between increased mental workload and heightened levels of mental fatigue among nurses, emphasizing the impact of demanding tasks in healthcare settings [61]. Similarly, research investigating the dynamics between mental workload, occupational fatigue, and physiological responses has revealed that heightened mental workload exacerbates occupational fatigue, thereby influencing well-being and work efficiency [62]. These findings highlight the critical need to monitor and regulate mental workload to mitigate mental fatigue and sustain high performance levels.

2.3.2 Characteristics of Mental Workload and Mental Fatigue in Physiological Signals

Numerous studies have investigated the characteristics of mental workload and mental fatigue. They have consistently demonstrated distinct patterns in physiological signals and eye-tracking data as individuals transition from non-fatigued states to various levels of fatigue. These findings highlight the potential for real-time monitoring and adaptive systems that can respond to changes in mental workload and mental fatigue, enhancing performance and safety in various applications. EEG is one of the most commonly used modalities in physiological signals for detecting mental workload. The meta-analysis of 24 studies demonstrated variability in EEG patterns and brain activity across different EEG regions, including the frontal, central, parietal, and occipital areas [63, 64]. Furthermore, correlations between mental workload and various frequency bands were observed, such as theta [4-8 Hz], alpha [8-13 Hz], low alpha [8-10 Hz], high alpha [10-13 Hz], beta [13-30 Hz], low beta [13-20 Hz], and high beta [20-30 Hz]. However, conflicting findings were noted in Table 0. For instance, Jensen et al. (2002) and Klimesch (2012) reported increased alpha power with elevated workload [65, 66], while other studies reported decreases [67, 68]. Similar contradictions were observed in studies examining the beta band, with some indicating increased power during heightened load [69, 70] and others showing the opposite [71]. In the delta frequency band, inconsistent results were also noted between the frontal and parietal regions [72]. Contradictory findings extended to theta and alpha bands as well [73, 74, 75]. Nonetheless, these EEG studies consistently demonstrate that increased cognitive workload and mental fatigue correlate with changes in specific frequency bands.

Table 0. Contradictory results of power in EEG frequency bands

Characteristics	Contradictory observation 1	Contradictory observation 2
Delta (< 4 Hz)	Increased power in frontal region [72]	decreased power in parietal region [72]
Theta (4–7 Hz)	Increased power during high cognitive workload [73][75].	No direct contradictory interpretation. Instead, task-related theta activity may decline with increasing mental workload [74]
Alpha (8–13 Hz)	Increased alpha power during high workload [65][66][69]	Decreased alpha power during high workload [67][68][70][75]
Beta (13–30 Hz)	Increased beta power during high cognitive workload [71][76].	Decreased beta power during high workload [77].

PPG and Blood Volume Pressure (BVP) signals exhibit increased heart rate and reduced variability under high workload conditions, reflecting the body's stress response [78]. GSR shows heightened skin

conductance levels indicating increased sympathetic nervous system activity as workload or fatigue increases. The more pronounced spikes are shown in acute stress or high cognitive demand [79]. Eye-tracking data provide additional insights: pupil size generally enlarges as cognitive workload rises due to increased mental effort, although prolonged high workload can lead to a decrease in pupil size due to fatigue [80]. Saccadic movements become less frequent and more erratic. Fixation duration tends to increase, indicating difficulty in maintaining attention and processing information efficiently [81]. Fixation patterns also shift with longer fixations and fewer saccades, indicating decreased visual search efficiency in high fatigue states [82]. Increased blink rate and longer blink duration are further indicators of mental fatigue as the brain attempts to alleviate strain and rest [83]. These multimodal measures collectively enable a comprehensive understanding of cognitive workload and fatigue. It facilitates the development of robust models for real-time monitoring and adaptive human-computer interaction systems [84].

2.3.3 Requirements of Physiological Recording Systems for Detecting Mental Workload and Mental Fatigue

Physiological signals need to be collected over sufficient periods to capture meaningful variations associated with different cognitive workload levels and mental fatigue states. Continuous data collection typically ranges from within a minute to several minutes or even hours. It depends on the complexity and context of the tasks being studied. For simple cognitive tasks, an experiment duration of 10-20 minutes might be sufficient to observe significant changes in physiological signals [85]. For more complex, e.g., multitasking or high levels of mental effort, it may only take a few minutes to show physiological responses correlated with changes in cognitive workload and mental fatigue. In prolonged tasks, the need to observe cognitive workload over extended periods may require data collection sessions of around one hour to assess sustained cognitive workload and mental fatigue [86]. A short duration can be helpful for capturing transient states, providing insights into immediate reactions or short-term cognitive demands. For example, tasks that induce acute stress or require rapid decision-making can lead to noticeable changes in heart rate variability, skin conductance and EEG activity within minutes [87]. In contrast, longer durations are essential for observing sustained workload and fatigue, revealing patterns and trends over extended periods. For instance, continuous monitoring over several hours can identify periods of increased mental fatigue that might not be apparent in shorter sessions during long driving sessions [88]. Existing systems of recording physiological signals (EEG, PPG, GSR and eye-tracking) possess the necessary specifications such as extensive duration and high resolution of data recording for extracting features reflecting user cognitive loads [89, 90, 91, 92]. They allow researchers and practitioners to monitor cognitive and fatigue states dynamically in real-time for prolonged durations.

2.4 User Preference for Self-Driving Behaviours

2.4.1 Individual Perception

Perception can be shaped by experience, trust in new technology, ability to understand nearby traffic situations, driving skills and age differences [93, 94, 95, 96, 97]. These factors can reflect individual variabilities, which influence preferences for self-driving technology. The impacts of this variability can manifest differently based on the user perceptions of AV along with the frequency and context of self-driving usage. Moreover, perception of nearby traffic situations can be strongly associated with user decisions to use self-driving features. Thus, it is essential to explore drivers' responses to various traffic situations and how their responses can be quantified to reflect users' perceptions, including the perceived difficulty levels of driving in each situation. However, this area has not been extensively studied with regard to multiple factors in traffic situations. Instead, previous research has examined users' perceived positive or passive impressions across different situations and found that preferences could vary even when the level of impression or passivity was similar in those situations [98]. This indicates that individual perceptions and the type of situation can lead to inconsistencies in how situations are perceived. Situation assessment involves achieving situation awareness which includes perceiving various elements within a situation, understanding the situation and predicting the future outcomes of current conditions [99, 100]. Individual preferences may depend on either the assessment of current traffic or the prediction of potential risks, which vary according to the situation's complexity, traffic conditions, and the individual's inherent perception.

2.4.2 Varied Driver Preferences with Multiple Factors

Studies have examined user preferences for self-driving cars in relation to various factors such as traffic situations [98, 101], demographic information [93], trust and acceptance of self-driving technology [102] and experiences with autonomous vehicles [103, 104, 105]. For example, in a study focused on AV user preferences in different self-parking scenarios, users preferred to send their vehicles to park at home if they had to pay for parking in a large metropolitan area [106]. Other studies have explored how self-driving cars could navigate parking in both indoor and outdoor environments [107, 108]. Research has also compared the preferences of novice versus experienced drivers for autonomous driving in different simulated environments, including expressways, urban areas, suburban and residential areas and parking lots [109]. The study found that experienced drivers generally favoured conventional driving due to its flexibility and driving pleasure whereas novices preferred autonomous driving for its ease and safety. However, both groups showed a preference for using a mix of autonomous and conventional driving depending on road and traffic conditions. Park et al. (2020) investigated driver preferences and perceptions of self-driving

functions in real-world driving conditions [98]. They highlighted that individual interpretations of situations could vary significantly, leading to variability in perceptions of AV decision-making. This suggests that individual differences and situational perceptions are crucial for understanding driver preferences for self-driving. More research is needed to encompass a broader range of traffic-related factors, including situation types, traffic complexity, and regional differences, especially in real driving contexts.

2.4.3 Cognitive Processing of User Preference

User preferences and cognitive processes are intricately associated, influencing each other across various domains. Research for cognitive psychology models shows that user preferences for mobile network services are shaped by different contexts emphasizing the importance of distributed cognition and levels of processing [110]. Similarly, Brain-Computer Interface (BCI) technology provides insights into user preferences by analyzing attentional responses and affective experiences related to visual content, highlighting the connection between cognition and emotional responses [111]. Cognitive styles significantly influence decision-making preferences. Holistic cognitive styles are more inclined toward goal-oriented explanations, while analytic styles prefer belief-based explanations, highlighting the impact of cognitive styles on user preferences [112]. The Cognitive Style Indicator (CoSI) further predicts individual differences in intuitive and deliberative decision-making processes that illustrate the influence of cognitive styles on decision-making [113]. In terms of representation preferences, users consistently favour diagrams over other formats in process models, highlighting the importance of visual representation for understanding complex information [114]. For older adults, engagement with mobile cognitive assessment and training programs is driven by factors such as engagement switches, cooperative interactions and overcoming technology barriers that enhance adherence to cognitive training. This emphasizes the need for user-centred design in technology tailored for older adults [115]. In e-learning environments, design elements such as isolation and distinctiveness promote memory retention and readability that influence user preferences for online learning systems [116]. Research on mobile phone preferences among Chinese college students, particularly males, shows an implicit preference for foreign brands over domestic ones, as demonstrated by Implicit Association Tests [117]. Similarly, older adults' preferences for walking spaces are shaped by multiple factors such as a complex interplay of cognitive and sensory needs, social interactions and the level of urbanity in their residence [118]. Adolescents develop preferences through choice deliberation, with value refinement becoming more pronounced over time. However, individuals with depressive symptoms show greater value refinement but lower value certainty and choice confidence [119]. Additionally, user preferences for video content significantly affect their Quality of Experience (QoE) during subjective evaluations, with users who have specific interests being

more critical in their ratings [120]. In summary, user preferences are deeply intertwined with cognitive processes and understanding these associations is essential for designing effective user-centred technologies and services.

2.4.4 Key Associations Between User Preference and Cognitive Process

Research combining eye-tracking and EEG data has demonstrated significant associations between user preferences and cognitive processes. These studies show that users exhibit distinct patterns of brain activity and gaze behaviour when evaluating preferred versus non-preferred items, suggesting that cognitive processing varies with user preferences. Specifically, users transition from an exploration mode when they spend more time examining non-preferred items to an exploitation mode focusing on preferred items [121]. In video summarization, eye-tracking and EEG signals have been used to develop a human visual attention-based framework. This framework leverages the relationship between gaze patterns and brain activity to identify key video segments that align with user preferences, enhancing the relevance and efficiency of video summaries [122]. A comprehensive review of user research utilizing EEG and eye-tracking has emphasized the importance of integrating these technologies to understand user preferences and cognitive processes. This integration provides insights into user interactions with various stimuli, revealing significant correlations between physiological responses and user preferences [123]. EEG and eye-tracking analyses have also been employed in automated driving tasks to assess situation awareness. These studies illustrate how cognitive processes and user preferences influence attention and decision-making in complex driving scenarios, with physiological signals being used to evaluate driver engagement and awareness [124]. Additionally, there is a study that assesses cognitive workload using EEG during tasks. It has been found that brain activity correlates with user preferences for task difficulty and interface design, suggesting that cognitive load experienced during interactions can reflect user preferences [125, 126, 127]. In the field of cybersecurity, EEG and eye-tracking have been used to study user responses to phishing attempts and malware warnings. The findings indicate that cognitive processes underlying user preferences for certain security measures can be detected through physiological signals that provide a foundation for improving user interface design for better security [128]. Collectively, these studies highlight the strong associations between user preferences and cognitive processes as measured by EEG, eye-tracking and other physiological signals that offer valuable insights for enhancing user experience through personalized and adaptive systems.

2.4.5 Requirements of Physiological Recording Systems in Perceiving User Preference for Self-Driving Behaviours

User preferences can be shaped by initial judgments such as first impressions and may evolve into post-preferences due to subsequent stimuli, environmental changes, or varying contexts. Given the dynamic nature of these preferences, careful epoch extraction is crucial when modelling changes using physiological signals and eye-tracking data. It is important to allow users sufficient time to form an initial preference judgment. Research suggests that forming a first impression typically takes between 100 milliseconds and several seconds, depending on the complexity of the stimuli and context [129, 130]. During this period, physiological signals and eye-tracking data should be collected to extract relevant features that reflect the user's initial preference formation. In traffic situations, the time required to understand various scenarios and make informed decisions is critical. Studies indicate that drivers need approximately 1.5 to 2.5 seconds to perceive and react to common traffic situations, such as sudden stops or lane changes [131, 132]. More complex situations, e.g., navigating intersections or responding to unexpected obstacles, may require longer response times [131, 132]. Continuous data collection during these periods is essential to accurately capture drivers' physiological and visual responses. After an initial judgment, user preferences can change due to additional information, emotional states, or contextual shifts. It is essential to provide sufficient time for users to perceive these changes and adjust their preferences. In traffic contexts, decisions such as adjusting speed and braking in response to new information can take several seconds to much longer depending on the situation's complexity, the cognitive load involved and age [133, 134, 135, 136]. For example, drivers can need over 10 seconds to reassess and take back control from automation [134]. Understanding these timeframes is essential for developing adaptive systems that can effectively support drivers' preferences. Current human physiological recording and eye-tracking technologies are capable of capturing physiological changes and gaze patterns in these contexts. Specifically, GSR provides insights into skin conductance over longer durations, while heart rate (HR) and heart rate variability (HRV) offer faster indicators of physiological changes and cognitive processing. However, the immediacy of these signals varies: EEG and pupil responses offer quicker detection of changes in user states which is typically within 200-300 milliseconds after stimulus onset than GSR and HR-based measures [137]. Consequently, GSR may not detect physiological changes if only minor, sudden changes occur in traffic situations. Conversely, EEG and pupil responses can capture reactions to these sudden changes. However, changes in these modalities may appear only briefly and may not be sufficient to predict user preference decisions during longer recording durations. Therefore, developing a preference perception model that adapts to different traffic situations is necessary. Understanding user preferences and their dynamic changes requires careful consideration of the time needed for both initial preference formation and subsequent decision-making

influenced by various factors. By properly segmenting epochs of physiological signals and eye-tracking data, more accurate models of user preferences can be developed. Furthermore, by adjusting epoch segments to detect preference changes in different traffic situations, an adaptive preference perception model can significantly increase accuracy in understanding user preferences across various traffic scenarios.

2.5 Research Questions

As discussed earlier, this thesis explores whether physiological and eye-tracking signals can be utilized to build AI models for granular classifications beyond the traditional binary classification models. Specifically, the primary goal is to investigate research topics such as error/potential risk detection, cognitive workload estimation, and user preference prediction. A set of detailed research questions has been formulated to achieve this goal, addressing different aspects of the models. Since physiological and eye-tracking signals are central to all studies in this thesis, it is essential to address issues related to artifact removal in these signals first. The corresponding research questions are as follows:

RQ1: What methods can be used to remove artifacts in physiological and eye-tracking signals, especially artifacts in EEG signals caused by eye movements?

RQ2: Are there advanced artifact removal methods, and do they outperform conventional approaches?

Next, the research focuses on developing individual user perception models. For the first model, potential risk detection, the research questions are:

RQ3: Can potential risk detection models be built using physiological and eye-tracking signals?

RQ4: If so, can these models incorporate more granular classes representing various risk conditions, such as different risk types or risk severity levels, instead of binary classifications (risky vs. non-risky)?

RQ5: Does model performance decline as granularity increases?

For the second model, cognitive workload estimation, the research questions are as follows:

RQ6: Can cognitive workload estimation models be built without relying on behavioural features that indicate severe drowsiness, such as blinking duration, eyelid closure rate, and head nodding?

RQ7: Can these models classify cognitive workload into finer-grained levels instead of binary classes if feasible?

RQ8: Similar to research question 5, how does increased granularity affect model performance for cognitive workload estimation?

RQ9: How do physiological and eye-tracking signals contribute to the accuracy of cognitive workload estimation models?

Lastly, the research for user preference studies in self-driving contexts is divided into two parts. The first part investigates user perception and preference differences through surveys, addressing the following questions:

RQ10: How does traffic complexity affect preferences for self-driving?

RQ11: How do individual perceptions of difficulty in each traffic situation and traffic speed influence preferences for self-driving?

RQ12: What is the association between preferred modes and preferred actions?

RQ13: How are preferred actions related to perceived difficulty levels?

RQ14: How does age influence user preferences for self-driving modes and actions?

RQ15: How does age affect perceptions (difficulty levels) of various traffic situations?

RQ16: How do age and traffic volume jointly influence user preferences and perceptions?

The second part focuses on developing user preference models for self-driving features and actions, with research questions including:

RQ17: Can user preference models be built using survey questionnaire data, and do the models' performances differ between age groups?

RQ18: Can user preference models be developed using physiological and eye-tracking signals?

RQ19: Does the performance of user preference models decline as the number of self-driving action options increases, leading to higher granularity and complexity?

RQ20: What are the performance differences between survey-based user preference models and those based on physiological and eye-tracking signals?

Addressing these research questions enables the development of various user perception models and highlights the need for an integrated framework to process physiological and eye-tracking signals efficiently. While this thesis does not test such a framework, it proposes a structure and direction for its potential implementation.

2.6 Toward a Meta-Framework by Integrating User Perception Models

The user perception discussed earlier can be modeled independently, yet their integration into a meta-framework has the potential to elevate their impact. Within a meta-framework, these models can dynamically interact based on the user’s physiological state, environmental conditions, or contextual information, offering a unified and adaptive system. This integrated system could revolutionize semi-autonomous vehicle interactions by delivering enhanced safety, personalization, and adaptability. For instance, a fully integrated meta-framework could proactively detect high-risk scenarios by combining environmental perception, such as high traffic speed or risky driving styles of nearby vehicles, with user physiological data to estimate or indirectly predict mental fatigue. At the same time, workload estimation and user preference models could fine-tune driver-vehicle interactions and decision-making processes, ensuring safer and more comfortable driving behaviours tailored to individual users. This concept aligns with findings in existing research, such as a cognitive modelling account of simultaneous learning and fatigue effects, which attributes increased errors to accumulating fatigue and reduced response times to learning [138]. Similarly, modelling the relationship between user states and user perception fosters trust and usability, addressing critical barriers to the widespread adoption of semi-autonomous vehicles. Moreover, a meta-framework offers scalability and flexibility, allowing new models or modalities to be seamlessly incorporated as technologies evolve. For example, additional sensors, e.g., LiDAR or advanced imaging systems, could be integrated to enhance decision-making in complex traffic scenarios while data sources from human physiological responses, e.g., EEG, GSR, PPG, etc, could provide deeper insights into user states. Moreover, a meta-framework offers scalability and flexibility, enabling the seamless integration of new models or modalities as technologies evolve. For example, additional sensors such as LiDAR or advanced imaging systems could enhance decision-making in complex traffic scenarios, while data from human physiological responses (e.g., EEG, GSR, PPG) could provide deeper insights into user states. This scalability extends beyond comparing task performance in dual-task experiments, such as the influence of cell phone distraction on driving performance [139], to scaling user perception models to broader applications. This modular approach ensures that the system remains relevant, adaptable, and capable of addressing the increasing complexity of human-vehicle interactions. Leveraging the strengths of meta-framework and learning such as additional insights relevant to environment contexts, user states, or tasks, future studies can push the boundaries of what intelligent vehicle systems can achieve. This integrated approach not only addresses the limitations of standalone perception models but also paves the way for a new generation of human-centric vehicle technologies. These advancements would bring us closer to the vision of vehicles that are not only autonomous but also perceptive, interactive, and responsive to the nuanced needs of their users. In summary, while this thesis primarily focuses on the development of individual user perception models, the potential for integrating these models into a unified meta-framework

is briefly explored. Meta-framework could bridge the gap between isolated models and fully adaptive systems, enabling vehicles to intelligently respond to diverse and evolving user and environmental conditions. This paradigm marks a vital step toward the realization of adaptive, user-centred, semi-autonomous systems that align with real-world complexities.

Chapter 3. Methodology

In this chapter, we present the general methodologies utilized in Chapters 4 through 6. Detailed descriptions of the methodologies applied within each of these chapters are provided within the respective sections.

3.1 Overview of Chapters

This chapter encompasses a range of distinct topics. Chapter 4 focuses on methods for artifact removal from various physiological signals, and these methods are consistently applied throughout Chapters 5 through 7 in this thesis. The studies presented in Chapters 5 through 7 are mainly centred around advanced vehicle assistance systems and user-centred design. Each is designed to address different specific research objectives. The target classification for AI-driven user perception models varies across these chapters: Chapter 5 aims to develop models for risk detection, Chapter 6 focuses on models for predicting cognitive workload, and Chapter 7 is dedicated to predicting user preferences. Each chapter features its research design, including traffic scenarios, survey designs, trial durations, experimental procedures (data collection methods), etc. Consequently, details on research design and participant information are provided within the respective chapters.

3.2 Recording Systems for Physiological Signals

For physiological signal recording systems, EEG, eye-tracking and other recording systems, such as PPG, EDA, Temperature, etc., are utilized in all studies in this thesis. Research ethics in each study are independently reviewed and approved by the Office of Research Ethics at the University of Waterloo. For an EEG recording system, we utilized the Enobio 32 system that accompanies wet EEG electrodes. The Enobio 32 system records 32 channels of EEG signals with 500 Hz. Other specifications are shown in Table 1. For eye-tracking data, we used the Ergoneers Dikablis 3 eye tracker. The eye-tracking system is a wearable wired binocular eye-tracker with 60 Hz eye camera tracking frequency, 0.05° visual angle pupil tracking accuracy and 0.1° – 0.3° visual angle glance direction accuracy. The eye camera resolution is up to 648×488 pixels, and the field (scene) camera resolution is 1920×1080 pixels with 30 fps (frames per second). All specifications are included in Table 2. The Empatica E4 system is equipped with four different sensors: PPG, EDA, digital temperature sensor, and accelerometer. The PPG sensor has a channel instead of multiple channels. Other specifications are shown in Table 3.

Table 1. Enobio 32 technical specifications [140]

Hardware	Specifications
Number of channels	32
Bandwidth	0 ~ 125 Hz (DC coupled)
Sampling rate	500 samples/sec
Dynamic range	24 bits – 0.05 (μ V)
Measurement noise	< 1 μ V RMS
Input impedance	> 1 G Ω
CMRR	-115 dB
Communication	USB
3 axes accelerometer	Yes (100 samples/sec)
Device dimension	89.1 mm x 61.1mm x 23.8mm
Operating time	19.0 hours

Table 2. Ergoneers Dikablis 3 Eye-tracker technical specifications [141]

Hardware	Specifications
The number of cameras	Field view camera (1) and two infrared cameras for pupil tracking
Sampling rate	60 Hz
Weight (complete device)	52 g
Frame dimensions	164.5 x 198.0 x 58.0 mm
Adjusting range field-cam	16° up and 30° down
Adjusting range eye-cams	Dual pivot adjustment for more flexibility and/or if a subject wears glasses
Field camera resolution	Full HD 1920 (H) x 1080 (V) @ 30 fps
Lens parameter	M9x0,5; IR Coating: T=50%@640±8nm FOV: Horizontal 116°; Vertical 60°; Diagonal 144,5°
Eye camera	648 x 488 (EIA & CCIR)
Image Transfer Rate	60 field/sec
IR Frequency	875 nm
IR intensity	17 mill-watts/steradian

Table 3. Empatica E4 technical specifications [142]

Hardware	Specifications
Sampling rate	PPG: 64 Hz; EDA: 4 Hz; Temperature: 4Hz; ACC (XYZ): 32Hz
Ranges	EDA: 0.01mSiemens ~ 100 mSiemens; Temp.: -40~115 °C (for skin temperature) and -40~85 °C (for ambient temperature); 3-axes ACC.: $\pm 2g$
Resolution	PPG output: 0.9 nW/digit; EDA: 1 digit ~ 900 pSiemens. Temperature: 0.02°C; 3-axes ACC.: 8 bits of the selected range
LEDs	Green (2 LEDs), Red (2 LEDs)
Electrodes	Semi-permanent (screw-in design); SUS03 stainless steel (standard) or Silver (Ag) plated with a metallic core
Motion artifact removal algorithms	Combines different light wavelengths; Tolerates external lighting conditions
Dimensions	Case: 44 x 40 x 16 mm; Wrist: 110 ~ 190 mm
Weight	25 g
Data transfer	Bluetooth / USB 2.0
Flash memory	Up to 60 hours of data storage
Battery	Lithium battery, 3.7V, 260 mAh capacity, > 36 hours in recording; > 20 hours in streaming mode

3.3 List of ML models

A total of 30 ML models were trained and tested in order to find the best-performing model. Table 4 lists the 30 ML models tested through the Python package called Lazy Predict [143]. Brief explanation of each model is included in Appendix A: Lazy Predict: 30 ML Models. From these, the top 2 models were chosen for each chapter to construct the classification models specific to the respective topics.

Table 4. A package of 30 ML models for classification performance testing in Lazy Predict [143]

30 ML models in the Lazy Predict Package	
AdaBoost Classifier	Nearest Centroid
Bagging Classifier	Passive Aggressive Classifier
Bernoulli Naive Bayes	Perceptron
Calibrated Classifier CV	Quadratic Discriminant Analysis
Decision Tree Classifier	Random Forest Classifier

Dummy Classifier	Ridge Classifier
Extra Trees Classifier	Ridge Classifier CV
Gradient Boosting Classifier	SGD Classifier
K-Nearest Neighbors	SVC (Support Vector Classifier)
Label Propagation	Extra Tree Classifier
Label Spreading	Gaussian Naive Bayes
Linear Discriminant Analysis	Gaussian Process Classifier
Linear SVC (Support Vector Classifier)	Hist Gradient Boosting Classifier
Logistic Regression	LightGBM Classifier
MLP Classifier (Multi-layer Perceptron)	XGBoost Classifier

3.4 Features Extracted from Physiological Signals

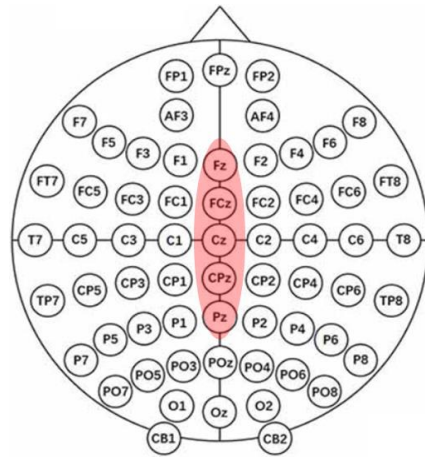
Throughout this thesis, features for various user perception models are extracted from physiological signals and survey data. The features from the survey data are only a part of the user preference study. Therefore, information regarding the survey data features is included in Chapter 7. The physiological features utilized in this thesis are extracted from EEG, eye-tracking, PPG, EDA and Temperature. First of all, there are two conditions for extracting features from EEG signals such as selecting EEG channels and frequency bands.

3.4.1 EEG Channels and Various Frequency Bands in EEG Signals

As our EEG system has a total of 32 channels, using all channels can cause excessive computation power. Existing research has shown that particular EEG channels are selectively analyzed as different tasks activate different sources in the human brain [144]. Error processing and potential risk perception are closely related cognitive processes that both involve our brain's ability to detect, evaluate and respond to potentially adverse or unexpected situations. For example, when an error occurs, such as a mistake in decision-making or a failure to respond appropriately to a stimulus or a situation, our brain can interpret this as an erroneous situation and possibly a potential risk depending on error types and the situations. The Error-Related Negativity (ERN) observed in EEG, particularly at channels like Cz and Fz, reflects this rapid recognition of an error. This immediate detection is crucial because errors often precede or coincide with risky situations that require corrective action to prevent adverse outcomes. For example, in a driving scenario, noticing either pedestrians or stopped vehicles in driving lanes may increase the perceived risk of an accident. The

cognitive error processing mechanism, therefore, serves as a warning system that helps identify and manage risks. In lots of existing research, brain potentials from error processing are clearly captured at the midline central region (Cz channel) and other channels around Cz such as Fz, FCz, CPz, and Pz [145, 146, 147], as shown in Figure 2. Furthermore, existing research pointed out that the fronto-central to central regions (FCz and Cz) are highly associated with contextual and information processing [148, 149]. Mirenowicz and Schultz (1994, 1996) and Montague et al. (1996) implied that the inner process of the midbrain region showed evidence of brain responses to salient events unpredicted. These channels are highly associated with user perceptions regarding detecting cognitive workload and user preference [150, 151, 152]. Therefore, we selected channels to extract EEG features throughout our experiments in the thesis.

Figure 2. Selected EEG channels in the existing studies.



In Table 5, there are four frequency bands for EEG analysis. Each of these frequency bands is associated with different physiological and cognitive states. For example, delta waves (0.5-4 Hz) are associated with unconscious processes that can be essential for understanding how the brain processes low arousal or passive states, such as those that might occur in a low-risk, monotonous driving environment in a simulator study. Theta wave (4-8 Hz) is particularly relevant in tasks that involve sustained attention and working memory. In the context of risk detection, changes in the power of the wave may be correlated with cognitive resources as drivers or users respond to potential threats on the road, e.g., when they are exposed to unexpected or novel stimuli which could signify heightened alertness and cognitive processing during higher-risk situations [153, 154]. Both alpha (8-12 Hz) and low-frequency (0.5-10 Hz) bands can represent the readiness and responsiveness of drivers or users to potential risks in a simulator study as well. For instance, variabilities in alpha and theta waves can also be associated with the physiological states of users

or drivers becoming more vigilant and cognitively engaged in emerging threats. Therefore, all these frequency bands are highly correlated with responses to risk events in various traffic scenarios.

Table 5. EEG frequencies used across this thesis

Frequency band	Delta	Theta	Alpha	Low
Freq. range [Hz]	0.5-4	4-8	8-12	0.5-10

3.4.2 Pupil Data from Eye-Tracking System

Existing research demonstrates that pupil data are closely associated with cognitive workload, potential risks and user preferences. In pupil data, we have three different data related to pupil size and movements: pupil size, pupil center shift degree and magnitude (PCD and PCM) shown in Table 6. For instance, changes in pupil size have been reliably correlated with variations in cognitive load during tasks and non-intrusive measures of mental effort [155]. Moreover, pupillary movements have been shown to correlate with mental workload in computer-based learning environments, further emphasizing their role in understanding user interactions with technology [156]. Studies have also highlighted that increased cognitive demands such as those encountered during complex navigational tasks, result in measurable changes in pupil size that are indicative of the user's perceived difficulty and associated risks [157]. The relationship between pupil size and cognitive workload in high-risk driving scenarios, such as during cell phone use in collision avoidance, further supports the relevance of these physiological responses in understanding user behaviour and preferences [158].

Table 6. Eye-tracking data used across this thesis

Data	Pupil size	Pupil Center shift Degree	Pupil Center shift Magnitude
Unit	Number of pixels	Degree	Pixels

3.4.3 Other Physiological Signals: PPG, EDA and Temperature

Other physiological data such as PPG, EDA and Temperature are increasingly recognized as indicators for detecting potential risks, cognitive workload and user preferences. PPG is sensitive to stress and provides insights into how individuals perceive and react to potential risks, highlighting its association with stress and work-related challenges [159, 160]. It may also reflect cognitive demands and user experiences, offering valuable information on how users interact with systems and how their preferences are altered by

these interactions [161, 162]. Similarly, EDA is closely associated with threat detection and emotional responses, making it a possible indicator for predicting cognitive workload and detecting user preferences [163, 164, 165]. Additionally, temperature changes have been identified as non-intrusive indicators of mental workload, and the changes can be correlated with user thermal preferences and cognitive ratings of thermal comfort level [166, 167, 168, 169]. Together, these physiological signals offer a comprehensive approach to assessing and enhancing user experience by aligning system design with users' cognitive and emotional states.

3.4.4 Extracted Features

Using all physiological data introduced above, lists of extracted features depend on types of physiological data, as shown in Table 7. EEG features are extracted from four frequency bands in each EEG channel. From eye-tracking and other physiological signals, there is a different list of features from the list of EEG features. These features have often been used in building user perception/state prediction models by extracting them from physiological signals [170, 171, 172, 173, 174, 175, 176, 177]. Details of extracted features in Table 7 are included in Appendix B: Extracted Features.

Table 7. List of features extracted from physiological signals

EEG features	Pupil features	Features (PPG, EDA and Skin temperature)
Mean energy	Power	Power
Log energy entropy	Relative power	Relative power
Mean	Mean energy	Mean energy
Mean (Relative signal)	Log energy entropy	Log energy entropy
Ratios of two bands	Shannon entropy	Shannon entropy
Ratio 1: A to B	Mean	Mean
Ratio 2: A to C	Variance	Variance
Ratio 3: A to D	Skewness	Skewness
The above features are extracted from each EEG channel (Cz, Pz and Fz)	Kurtosis	Kurtosis
	Hjorth activity	Hjorth activity
	Hjorth mobility	Hjorth mobility
	Hjorth complexity	Hjorth complexity
	Normalized 1st difference	Normalized 1st difference
	Normalized 2nd difference	Normalized 2nd difference

- A is one of the four frequency bands, and others (B, C, D) are the remaining bands.

These features were utilized to build user perception models in the studies presented in Chapters 5 through 7. In constructing these models, specific physiological responses are expected to be reflected in the features. For instance, mean energy and power can be associated with cognitive engagement in perceived risks, user comfort reflecting preferences and the demands of cognitive workload. Hjorth parameters are valuable for assessing the dynamics of signals and temporal changes in perceived risks and varying behaviours of self-driving systems. Additionally, other features such as variance, entropy and power are closely related to the complexity of cognitive processing, attention, perception of visual stimuli and stress levels. Therefore, comprehensively employing these features is crucial for developing highly accurate user perception models. Alongside model construction, we compared the performance of these perception models and conducted a feature importance analysis to determine the significance of each feature. Instead of interpreting each feature individually, we analyzed their importance based on the aggregated sum within each categorization, such as the sum of all features extracted from the Fz channel or the sum of all features extracted from the pupil center shift degree (as listed in Table X1). Prior to feature extraction and the application of ML models, signal preprocessing and artifact removal procedures are detailed in the following chapter.

Chapter 4. Artifact Removal of Physiological and Eye-Tracking Signals

Physiological signals are used in experimental studies in this thesis. These signals provide insights into human physiology and cognitive states, which can fluctuate in response to external stimuli over time. However, these signals are often contaminated by noise and artifacts that can significantly degrade the quality of the data and impair the accuracy of subsequent analyses. Noise and artifacts can arise from a multitude of sources, including environmental interference, movement artifacts and physiological processes that are unrelated to the signals of interest. For instance, EEG signals can be disrupted by electrical noise, muscle activity or eye movement. Blinks or head movements can distort eye-tracking data. Similarly, motion artifacts or sensor contact variations can affect PPG, leading to erroneous readings that do not accurately reflect the underlying physiological processes. The challenge of removing these unwanted components is foremost for ensuring the reliability and validity of biometric signal analysis. Effective denoising techniques are essential not only for improving signal quality but also for enhancing the interpretability and accuracy of the data. Various methods have been developed to address these challenges, ranging from simple filtering techniques to more advanced approaches such as independent component analysis (ICA), wavelet decomposition and machine learning-based algorithms. Table 8 shows the sampling rate of each signal modality. In the following sections, we introduce artifacts of each biometric signal type and suggest methods of removing artifacts of the signal.

Table 8. Sampling rates of physiological signals used across this thesis

Signal type	EEG	Eye-tracking	BVP	EDA	Temperature
Sampling rate [Hz]	500	60	64	4	4

4.1 Artifact Removal of EEG Signals

EEG is a powerful and non-invasive technique that monitors and analyzes brain activity, providing invaluable insights into neurological functions and cognitive processes. However, the utility of EEG is often compromised by the presence of artifacts unrelated to salience brain activity. These artifacts can originate from various sources, such as eye movements, muscle activity, cardiac signals, and environmental noise. The presence of such artifacts can obscure the actual brain responses in EEG signals and lead to misinterpretations and inaccurate analyses. Among these artifacts, those generated by eye movements are particularly problematic as the experiments in this thesis are designed to involve tasks requiring active search and visual inspection by participants. In this setup, continuous saccadic movements and blinks produce strong signals that can dominate the EEG recordings, potentially contaminating the main signals

related to the designed stimulation [178, 179]. Ocular-related artifacts, particularly electrooculographic (EOG) artifacts, are predominantly observed in the frontal regions [178, 179]. These artifacts are challenging to remove because they often overlap with brain signals in both the time and frequency domains. If not adequately addressed, EOG artifacts can significantly distort the EEG data, making it difficult to accurately interpret brain responses or cognitive states from stimulation. Artifact removal in EEG signals is thus a critical step in preprocessing aimed at enhancing the signal quality and ensuring the accuracy of subsequent analyses. Traditional approaches for artifact removal include simple filtering techniques that attempt to eliminate specific frequency bands associated with artifacts. Beyond the simple filtering techniques, other methods such as Common Average Reference (CAR), Independent Component Analysis (ICA), Wavelet Transform, and Blind Source Separation (BSS) have also shown promise in either effectively isolating and removing artifacts or reducing noises while preserving the integrity of the underlying neural signals [181, 182, 183]. In the following sections, we introduce two methods (CAR and ICA) and how these methods can be revised with eye-tracking data.

4.1.1 Two Common Algorithms for Filtering Artifacts in EEG Signals

4.1.1.1 Common Average Reference (CAR)

The Common Average Reference (CAR) algorithm is a widely used technique in EEG signal processing, particularly for enhancing the quality of recorded signals by mitigating the effects of common noise sources. The CAR algorithm is based on the assumption that specific noise components are uniformly distributed across all electrodes. By eliminating these common signals, CAR enhances the distinct neural activity recorded at each electrode, thereby improving the overall signal-to-noise ratio. The basic principle behind CAR involves averaging the signals from all EEG channels and subtracting this average from each individual channel. It can be mathematically represented as follows:

$$X_{CAR} = X_{raw} - \frac{\sum X_i}{N_{channels}}$$

where X_{raw} is the raw signal, X_{CAR} is the filtered signal by CAR. X_i is the signal of channel i in the raw signal, and N is the number of channels.

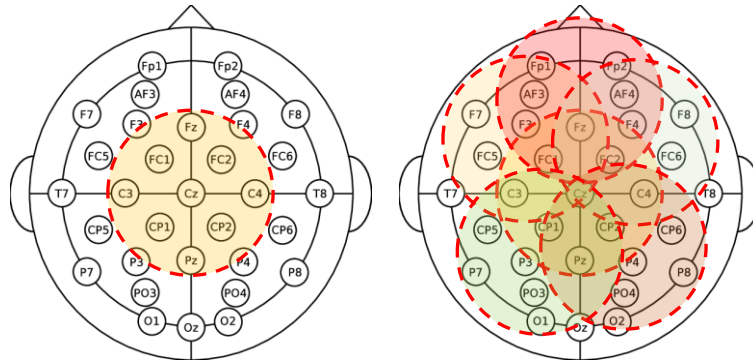
The CAR technique is often used in conjunction with other signal processing techniques to improve the signal quality, such as filtering or artifact removal. This process effectively removes the common signal

component shared by all electrodes, leaving behind the unique signal at each electrode reflecting localized brain activity. This makes CAR particularly powerful in scenarios where noise sources such as power line interference or general physiological noise affect all electrodes in a similar manner. Despite its effectiveness in reducing common noise, the CAR algorithm has certain limitations. In particular, it is not specifically designed to remove artifacts associated with eye movements, such as saccades and blinks that produce large and localized artifacts in the EEG data. These artifacts may not affect all electrodes equally and thus are not entirely eliminated by the CAR procedure. In order to address this issue, we suggest a revised method of the CAR algorithm.

4.1.1.2 Revised Common Average Reference (R-CAR)

The Revised Common Average Reference (CAR) Algorithm refines the traditional CAR approach by calculating the reference signal in saccadic movements and blinks using only a subset of electrodes that are geographically close to the target electrode. This approach can reflect a localized representation of the reference signal by focusing on nearby channels and minimizing the effects of localized artifacts by ocular activities, e.g., especially saccadic movements and blinks. By tailoring the reference in saccadic movements and blinking to the spatial context of each electrode with neighbouring channels, the R-CAR algorithm can selectively reduce noises that occur during ocular activities and depend on either brain regions or EEG channels. This approach can be particularly beneficial in high-density EEG recordings where the proximity of electrodes allows for precise selection of neighbouring channels. Figure 3 illustrates the selection of channels for re-referencing the Cz channel.

Figure 3. Selected channels for re-referencing of Cz channel (Left) and various sets of selected channels that each region represents a subset of channels for re-referencing of a channel (Right)



Using this approach (R-CAR), we separately extracted the average components related to saccadic and non-saccadic eye movements as identified through eye-tracking data and applied these corrections to the raw

EEG signals. The principle behind R-CAR involves two different re-referencing approaches for saccadic movement/blink and fixation durations, respectively. For fixation duration, the original CAR algorithm is used. For saccadic movement and blink durations, a localized approach of CAR is used as shown in Figure 4. These can be mathematically represented as follows:

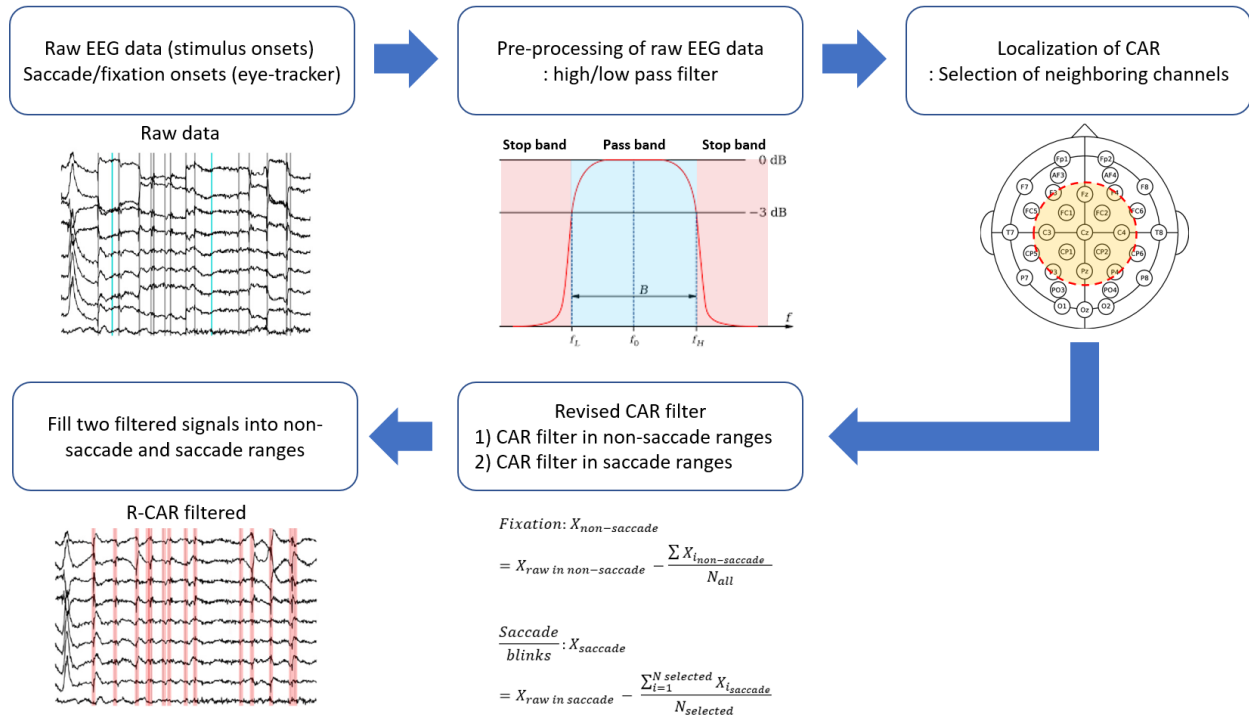
$$\text{Fixation: } X_{non-saccade} = X_{raw \text{ in non-saccade}} - \frac{\sum X_{i_{non-saccade}}}{N_{all}}$$

$$\text{Saccade/blinks: } X_{saccade} = X_{raw \text{ in saccade}} - \frac{\sum_{i=1}^{N_{selected}} X_{i_{saccade}}}{N_{selected}}$$

Where $X_{non-saccade}$ and $X_{saccade}$ are re-referenced signals in non-saccade (fixation) and saccade (saccadic movement/blink). N_{all} indicates all channels in the utilized EEG system and $N_{selected}$ indicates selected channels for re-referencing of each channel in the EEG system.

From the equations, these filtered signals, each corresponding to either the saccade or non-saccade (fixation) range, are reconstructed by replacing the raw signal during the respective saccade or non-saccade periods. The process of preprocessing EEG signals using R-CAR is shown in Figure 4.

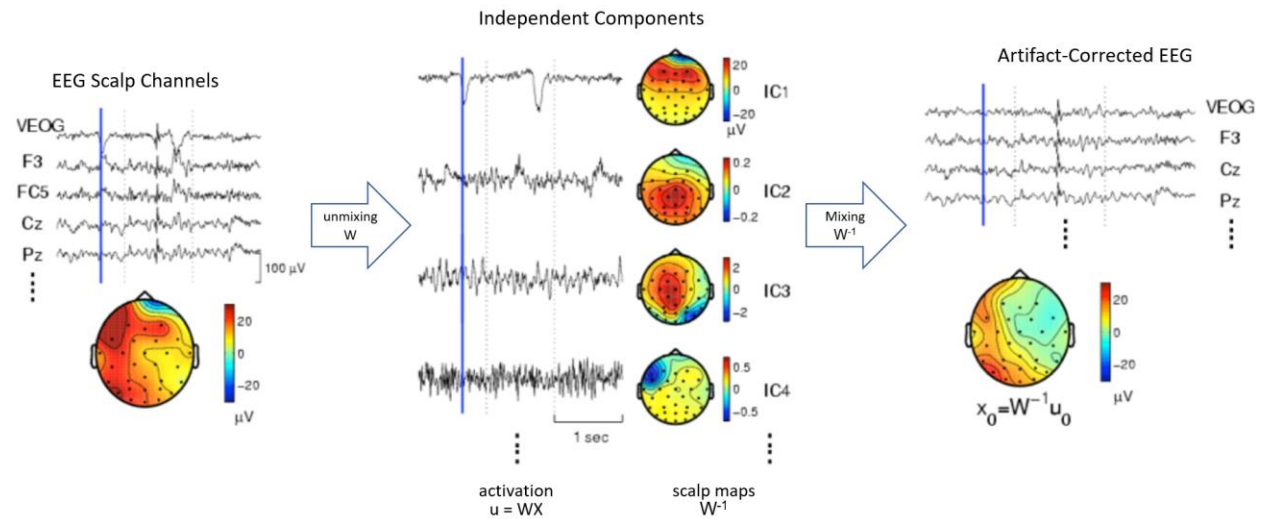
Figure 4. A schematic diagram of preprocessing with the R-CAR algorithm



4.1.1.3 Independent Component Analysis (ICA)

Independent Component Analysis (ICA) is a powerful technique that has widespread applications in EEG analysis. As multivariate time series data, EEG is complex and complicated in separating noise/artifact sources from raw EEG. The technique is particularly useful in separating independent components (ICs) of EEG signals, which can represent underlying brain processes and components affected by noise/artifacts. Each component represents a source of activity statistically independent of the other components by maximizing the non-Gaussianity of the components and minimizing the dependence between the components. Among the extracted ICs, it is required to select components that correspond to neural activity by various methods such as visual inspection of components, time-frequency analysis and comparison with Event-Related Potentials (ERPs). Mostly, a few researchers or experts can independently select neural components in these selection processes. The selected components can be reconstructed into a time series of primary brain activity. The schematic diagram is shown in Figure 5.

Figure 5. A schematic diagram of EEG artifact removal with blind source separation in Independent Component Analysis [184].



As shown in the diagram above, the ICA algorithm faces a challenge in automatically separating ocular components that be manually separated by experts or researchers. An additional algorithm can be incorporated to automatically separate ocular components within the ICA algorithm by using eye-tracking data to address this issue. The following section includes details on the automatic separation process of ICA and the overall preprocessing procedure.

4.1.1.4 Automatic Independent Components Analysis (A-ICA)

To automatically select ocular components from all independent components (ICs), we utilized the method suggested by other research [178]. This method requires eye-tracking data, especially fixation and saccade, to evaluate the variance of each independent component in both saccade and fixation epochs. For the variances measured separately in saccades and fixations, the mathematical representation is shown as follows:

$$Var_{epoch_j} = \frac{1}{N} \sum_i^N (y_i - EEG_{mean_j})^2$$

where i is i -th signal point in either saccade or fixation epoch and N is the number of the signal points in each fixation epoch for variance in fixation epoch. EEG_{mean_j} is the mean EEG in j -th fixation epoch, Var_{epoch_j} is the variance EEG in j -th fixation epoch.

For the variance calculation in a saccade epoch, it uses the same process as the process for the fixation epoch. Then, average variances in all fixation and saccade epochs are calculated, respectively. This can be mathematically represented as follows:

$$Var_{fixation \text{ or } saccade} = \frac{1}{M} \sum_j^M Var_{epoch_j}$$

where j is j -th signal point in either saccade or fixation epoch and M = the number of either saccade or fixation epochs. The calculated variances in fixations and saccades are used for the variance ratio as below.

$$Variance \text{ ratio} = Var_{saccade} / Var_{fixation} \quad (in \text{ a trial})$$

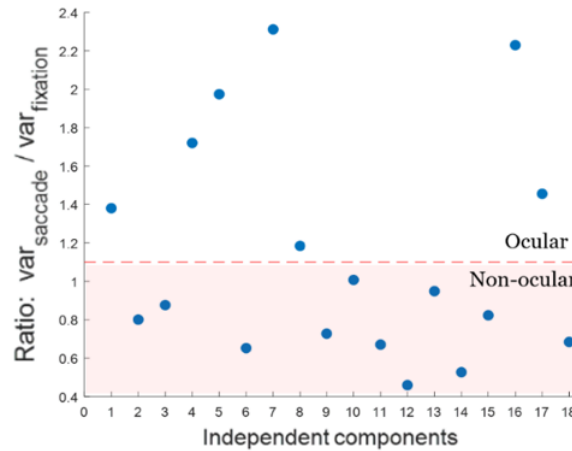
The calculated ratio for each IC serves as the evaluation criterion. ICs with greater variance during saccades likely indicate ocular artifacts, whereas those with greater variance during fixations are more likely to represent non-ocular artifacts. The ratio of variances, $Var_{saccade}/Var_{fixation}$, exceeds 1.0, suggesting that IC is closer to an artifact, whereas ICs with ratios below 1.0 are neural sources. However, a ratio value within 10% of 1.0 in ICs may not be suitable for distinguishing ocular artifact components from all ICs due to the high risk of misclassification. Our preliminary results recommend a ratio value between 1.05 and 1.1 to minimize misclassification issues which is close to the value recommended by Ploch et al. (2012) [178]. To compare ICs with EEG channels in the prefrontal area where ocular artifacts are highly pronounced. Regarding the selection of ocular components, three Ph.D. students with expertise in physiological signals

and signal processing were involved in the classification process. They selected around four components as ocular artifacts based on topographical maps and amplitudes of ICs. In the automated selection process using the ratio of variance in fixations and saccades, Figure 6 shows the ratio results of all ICs and the separation of ICs for ocular artifacts using the ratio value 1.1. The result shows six independent components highly associated with ocular artifacts over 1.2 and an independent component between 1.1 and 1.2. Table 9 shows the comparison of component selections by ICA and A-ICA. This indicates that the automated selection algorithm (A-ICA) can be more sensitive in sorting the components highly associated with ocular artifacts since the algorithm results in a total of 7 ICs.

Table 9. Average number of selected ocular components by ICA and A-ICA in 20 trials.

Algorithm type	Single stimulus experiment		Multi stimuli experiment	
	avg. N	SD. N	avg. N	SD. N
ICA (manual)	3.5	1.1	4.4	1.9
A-ICA (automatic)	6.5	2.8	6.8	3.1

Figure 6. Variance ratios of independent components



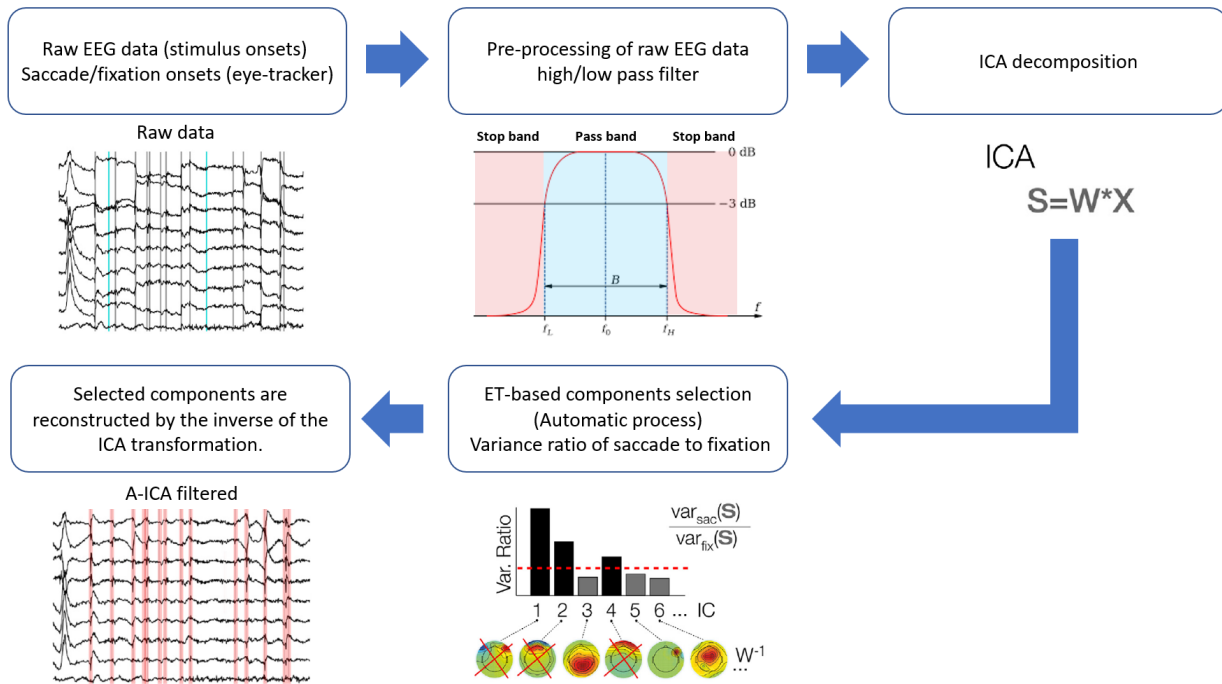
Including the automated selection algorithm, Figure 7 illustrates a potential preprocessing procedure for EEG signals. The first step involves applying high-pass and low-pass filters to the raw signals with filter parameters set to 2 Hz and 40 Hz, respectively, to remove unwanted noise and focus on the relevant frequency bands. Following the filtering, the data undergoes ICA decomposition where the EEG signals are decomposed into independent components. These components include both ocular (eye movement-related) and non-ocular components. Then, the ET-based (Eye tracking-based) component selection process

is employed. After removing ocular artifact components, other components are reconstructed by applying the inverse of the ICA transformation. This can be mathematically represented as follows:

$$X' = A \times S'$$

where S' is the modified matrix of ICs (e.g. with artifact components removed), and X' is the reconstructed EEG data.

Figure 7. A schematic diagram of preprocessing with the A-ICA algorithm



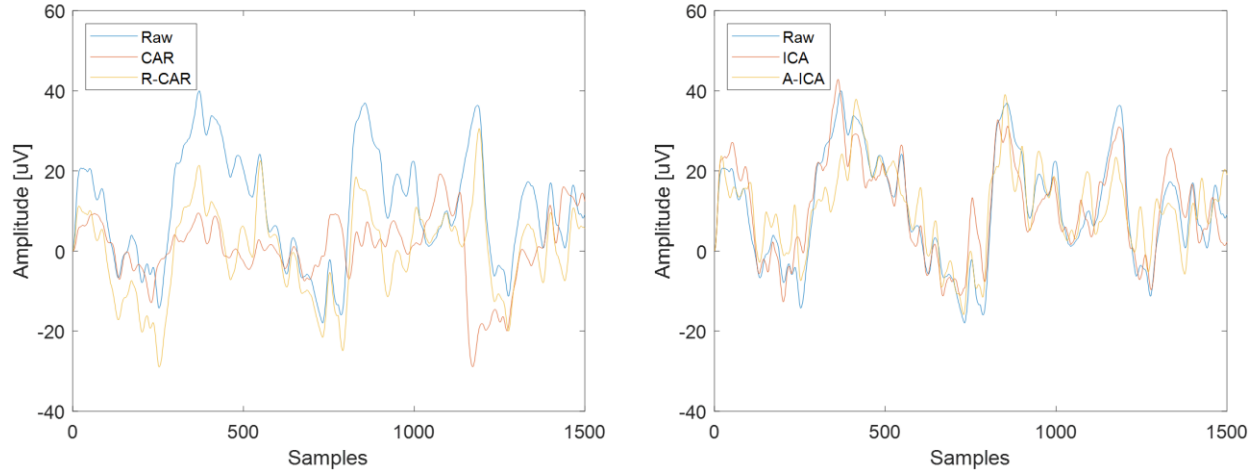
For a simplified algorithm, the process of spike potential removal in [178, 185] is excluded. Also, it is important to note that there may be challenges in synchronizing the data due to the difference in sampling rates between the eye-tracking system (60 Hz) and the EEG system (500 Hz).

4.1.2 Comparison of Preprocessed Signals

This section compares the filtered signals from the preprocessing procedures introduced in the previous section. Firstly, Figure 8 shows the filtered signals preprocessed by all procedures introduced above. Both CAR and R-CAR filters have a lower amplitude than the raw EEG signal because the CAR filters affect the overall signal. As a result of the R-CAR filter, the degree of attenuation/amplification is dependent on

saccade and fixation regions. Furthermore, the amplitude value is relatively lower than the raw signal but the overall shape or trend of the signal appears similar. ICA-based algorithms show minimal deviation from the raw EEG signal with the shape of the signal appearing similar. Compared to CAR-based algorithms, ICA-based algorithms have a higher degree of visual similarity to the raw signal as well as higher amplitudes than the raw signal in a few ranges of the time-series filtered signal.

Figure 8. EEG amplitudes from raw signals, CAR, R-CAR, ICA and A-ICA.



To look into more results from the preprocessing algorithms, we also compare signal amplitude variances in fixations and saccades between the preprocessing algorithms: raw, CAR, R-CAR and A-ICA (all band-pass filtered with the same band frequency range (0, 40) in Table 10). The average variance of raw signal in saccade is only a 7.60 % difference from the value in fixation which is smaller than the values in other preprocessing algorithms. The raw signal preprocessing in both fixation and saccade conditions has the highest average (SD) value. The CAR preprocessing showed that the average (SD) is around 35% (47%) lower than the values from the raw signal. However, the ratio of the average fixation to saccade, 0.924, is much lower than the ratio in the raw signal, 0.640. The ratio values from other conditions, R-CAR and A-ICA, are lower than the ratio from the CAR. The R-CAR improves the signal quality by largely decreasing the average (SD) values in saccade compared to the values of the CAR preprocessing. The average (SD) values in fixation are higher than the CAR values. The SD value of R-CAR in the saccade is lower than that of R-CAR in the fixation, while the CAR has the opposite trend of the SD values between the fixation and saccade. It implies significantly reducing the signal variance in saccade and limiting the reduction of the signal variance in fixation by R-CAR. In other words, R-CAR may not effectively remove noises in fixation since it still shows a higher SD than the SD values by CAR and A-ICA. The A-ICA algorithm showed the most similar difference of the SD values between fixation and saccade with the lowest average

variance in saccade. It turns out the lowest average variance ratio of the saccade to fixation among all preprocessing algorithms. As it significantly reduces the average variance in saccade, the signal power in saccade is no longer effective enough for ascendancy over the power in fixation. We may therefore come to the conclusion that the A-ICA outperforms other preprocessing to stand out the power in fixation compared to the power in saccade by minimizing EEG artifact from eye movement.

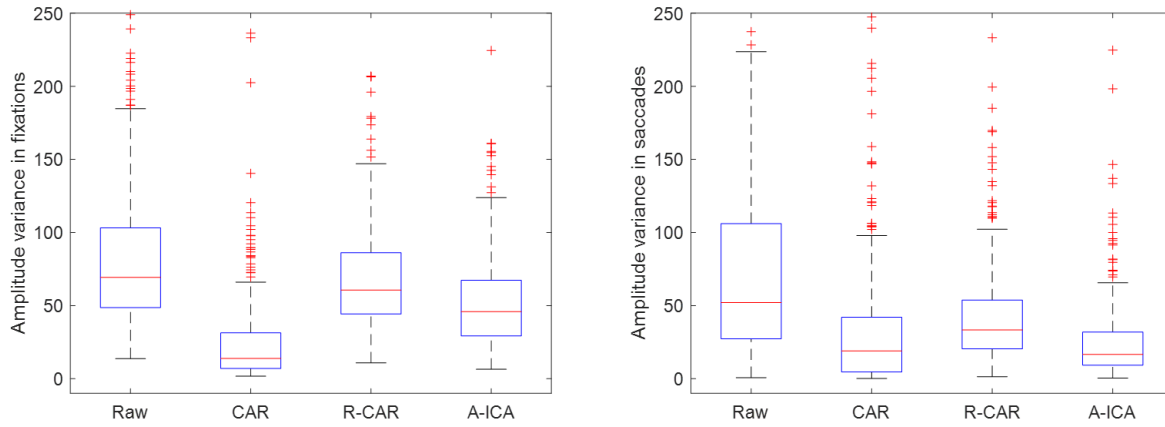
Table 10. Average and standard deviation of filtered EEG amplitude's variances from four different preprocessing conditions in fixations and saccades: Raw, CAR, R-CAR and A-ICA

Categorization	Raw	CAR	R-CAR	A-ICA
Fixation	83.8 (56.9)	55.0 (30.3)	71.6 (43.8)	53.0 (34.0)
Saccade	77.4 (75.7)	35.2 (51.7)	43.4 (34.1)	26.6 (37.0)
Ratio of Saccade/Fixation	0.924 (1.33)	0.640 (1.70)	0.606 (0.78)	0.498 (1.09)

- All preprocessing includes the same bandpass filter [0.05, 40] Hz.

We compare amplitude variance in fixations and saccades for further analysis using box-and-whisker plots. The result is shown in Figure 9 and Table 11. Similar to the previous analysis of average variances, the interquartile range ($IQR = Q3 - Q1$) is largely reduced by CAR in fixation in Figure 9. Compared to the raw signal, both R-CAR and A-ICA reduced the IQR but not as much as CAR. For both the IQR and the range between LIF and UIF (Lower Inner Fence: $LIF = Q1 - 1.5 * IQR$ and Upper Inner Fence: $UIF = Q3 + 1.5 * IQR$), CAR is ranked the lowest, followed by A-ICA, R-CAR and the raw signal is listed in fixation. For the IQR and the LIF-UIF range, A-ICA is ranked the lowest in saccades, indicating that A-ICA outperforms removing signal variance within ± 2.7 sigma. In all preprocessing, the box plots were positively skewed, and the number of outliers above UIF is 12~27 trials in both fixation and saccade. Similar to the result in the previous analysis, R-CAR and A-ICA perform in a way that largely reduces noise sources from saccade, while signal in fixation shows relatively low changes by noise reduction. Particularly, A-ICA performs better than other preprocessing methods in reducing noise while also exhibiting filtered signals with the most minimized IQR during saccades.

Figure 9. Filtered EEG amplitude's variances from raw, CAR, R-CAR and A-ICA algorithms with a bandpass filter over 300 trials: Left (variances in fixations) and Right boxplot (variances in saccades)



- LIF ($Q1 - IQR \cdot 1.5$) values for all conditions are mostly zeros or close to zeros.

Table 11. Values for the box plots from filtered EEG amplitude's variances from four different preprocessing conditions in fixations and saccades: Raw, CAR, R-CAR and A-ICA

Fixation (355 trials)				
	Raw	CAR	R-CAR	A-ICA
LIF or Min. value	13.7	1.7	10.8	6.5
Q1	48.7	7.0	44.3	29.3
Q2	69.3	13.9	60.6	45.9
Q3	103.2	31.4	86.2	67.3
UIF ($Q3 + IQR \cdot 1.5$)	185.0	67.9	149.1	124.3
IQR	54.5	24.4	41.9	38.0
# of outliers (over UIF)	19	27	14	12
Saccade (336 trials)				
	Raw	CAR	R-CAR	A-ICA
LIF or Min. value	0.8	0.2	1.4	0.5
Q1	27.4	4.7	20.5	9.3
Q2	52.1	19.0	33.3	16.6
Q3	106.1	42.0	53.8	32.0
UIF ($Q3 + IQR \cdot 1.5$)	224.2	98.0	103.8	66.0
IQR	78.7	37.3	33.3	22.7
# of outliers (over UIF)	17	27	20	22

4.2 Artifact Removal of Eye-Tracking Data

Artifacts in eye-tracking data typically arise from both physiological and technical sources. Physiological artifacts include blinks, saccades and head movements that cause sudden changes or interruptions in the recorded eye position. These events are natural but can create discontinuities or spikes in the data. On the technical side, noise can originate from the limitations of the eye-tracking equipment, such as calibration inaccuracies, reflections from glasses or eyelashes and fluctuations in lighting conditions. On the physiological side, blinks, saccades, and head movements naturally occur while experiments proceed. Artifacts from these sources can be removed by a preprocessing procedure. Our preprocessing procedure is introduced in the following sections.

4.2.1 Hampel Filter

In our preprocessing procedure, we employ the Hampel filter to remove artifacts from pupil data. The Hampel filter is a robust statistical technique used for outlier detection and noise reduction in time series data. It is particularly effective in identifying and correcting anomalous data points, such as those introduced by artifacts in eye-tracking data. The Hampel filter is based on the principle of median absolute deviation (MAD), a robust measure of statistical dispersion. The following steps explain the Hampel filter [186].

1) Window length and centering

The Hampel filter operates on a sliding window of data, where the window has an odd length (Length = $2k+1$). This window moves across the data sequence by centring on each sample. In the example provided, the window length is 5, meaning that k is 2.

2) Calculation of local median and MAD

For the current sample (the sample being analyzed), the algorithm computes the median of the data points within the window. It then calculates the Median Absolute Deviation (MAD) that is a robust measure of the spread or variability of the data within the window. The MAD is computed as:

$$m_i = \text{median}(x(i - k), x(i - k + 1), \dots, x(i + k))$$
$$MAD_i = \text{median}(|x_{i-k} - m_i|, \dots, |x_{i+k} - m_i|),$$

3) Outlier detection

The algorithm checks whether the current sample deviates significantly from the local median by comparing it to a threshold based on the MAD. The threshold is defined as:

$$\lambda \times \sigma_i$$

where λ is the threshold multiplier (typically set to a value like 2 or 3), and $\sigma_i = \kappa \times MAD_i$ is an estimate of the standard deviation derived from MAD, where κ is a constant approximately equal to 1.4826. If the absolute difference between the current sample and the median of the data points is larger than $\lambda \times \sigma_i$, then the current sample is identified as an outlier.

4) Outlier replacement

If the current sample is identified as an outlier, it is replaced with the local median calculated earlier. This correction ensures that the data remains smooth and that outliers are removed.

$$x_s(\text{corrected}) = m_i$$

5) Sliding the window

The window then slides one step forward, and the procedure is repeated for the next sample in the sequence. The algorithm continues this process until all samples in the data sequence have been processed.

Example Walkthrough:

- Consider a data sequence: [1,2,4,9,23,8,12,4,2].
- With a window length of 5, the first few windows include samples like [0, 0, 0, 1, 2], [0, 0, 1, 2, 4], etc., with zeros prepended to ensure a full window at the beginning of the sequence.
- When the window is centred at 9, the median might be calculated as 4, and the MAD is calculated accordingly. Since 23 is a significant deviation from the median (greater than the threshold), it is identified as an outlier and replaced with the median of the local window, resulting in a corrected sequence.

The final corrected sequence will have outliers replaced by median values, resulting in a smoother dataset free of sudden and erratic changes caused by outliers.

4.2.2 Preprocessing Procedure for Eye-Tracking Signals

To mitigate these technical artifacts, we monitored gaze-tracking trajectories in real-time during sessions and re-calibrated the system before each session if necessary. The experiments were conducted in a controlled lab environment with consistent lighting to minimize the impact of lighting variations. The wearable eye tracker used in this study is designed to be securely fastened to reduce the likelihood of artifacts due to head movements. In cases where head movements distorted gaze tracking, we corrected

these distortions through recalibration. However, artifacts resulting from head movements and technical sources can still appear in the pupil data alongside artifacts from saccadic movements and blinks. To address this, we decided to analyze data with these artifacts minimized as much as possible. We applied a preprocessing procedure as illustrated in Figure 10 to effectively remove these artifacts. The preprocessing begins with collecting raw eye-tracking signals with the onset of saccades and fixations. These saccades and fixations are detected under two criteria: velocity threshold and dispersion threshold [187]. Other spikes in the raw data are also detected based on the same thresholds for saccades and fixations. Although adjusting the threshold values in spike detection can increase sensitivity, it also raises the possibility of misclassifying important pupil data as saccades or other artifacts. Therefore, setting these values with reference to those used in other studies is vital. In our preprocessing procedure, the first step involves interpolating the coordinates between the start and end points of each saccade followed by smoothing the data to reduce spontaneous spikes. However, this is not sufficient to eliminate all spontaneous spikes, as some residual spikes may remain even after interpolation based on the thresholds set in this study. A Hampel filter was applied to address these residual spikes. The Hampel filter, a median-based method, is applied to identify and correct any remaining outliers in the data. Finally, the preprocessing procedure removes the saccadic coordinates and corrects the x and y coordinates of gaze derived from the pupil data. This step yields stabilized pupil data by replacing abrupt changes caused by saccades and blinks with predicted values. Additionally, this preprocessing procedure addresses other sources of artifacts that may introduce sudden changes in pupil data by substituting them with predicted values as well. This approach is particularly effective for analyzing the characteristics of pupil changes without the influence of artifacts, especially in cases where short fixations of around 100 milliseconds are observed continuously alongside multiple saccades resulting from artifacts. In the following figures, an example of the preprocessing procedure above is shown. In Figure 11, the left figure shows raw pupil data and spikes in the data are presented in orange regions. The right figure shows the data after the interpolation in spikes. There are still some scaled-down spikes left. To remove these remaining spikes, we applied the Hampel filter, and the filtered result is shown in Figure 12. Using these two steps of preprocessing, we were able to effectively remove sudden spikes in the pupil data. This preprocessed data is analyzed based on fixations, e.g., extracting epochs with a certain epoch duration from the start timing of each fixation. Other pupil data, such as 2D coordinates of pupil center changes, are also preprocessed using the same steps as the pupil center changes, which are also analyzed based on fixations. For fixation, saccade and blinks, these data are analyzed based on the threshold values previously introduced in Chapter xx.

Figure 10. A schematic diagram of preprocessing eye-tracking data

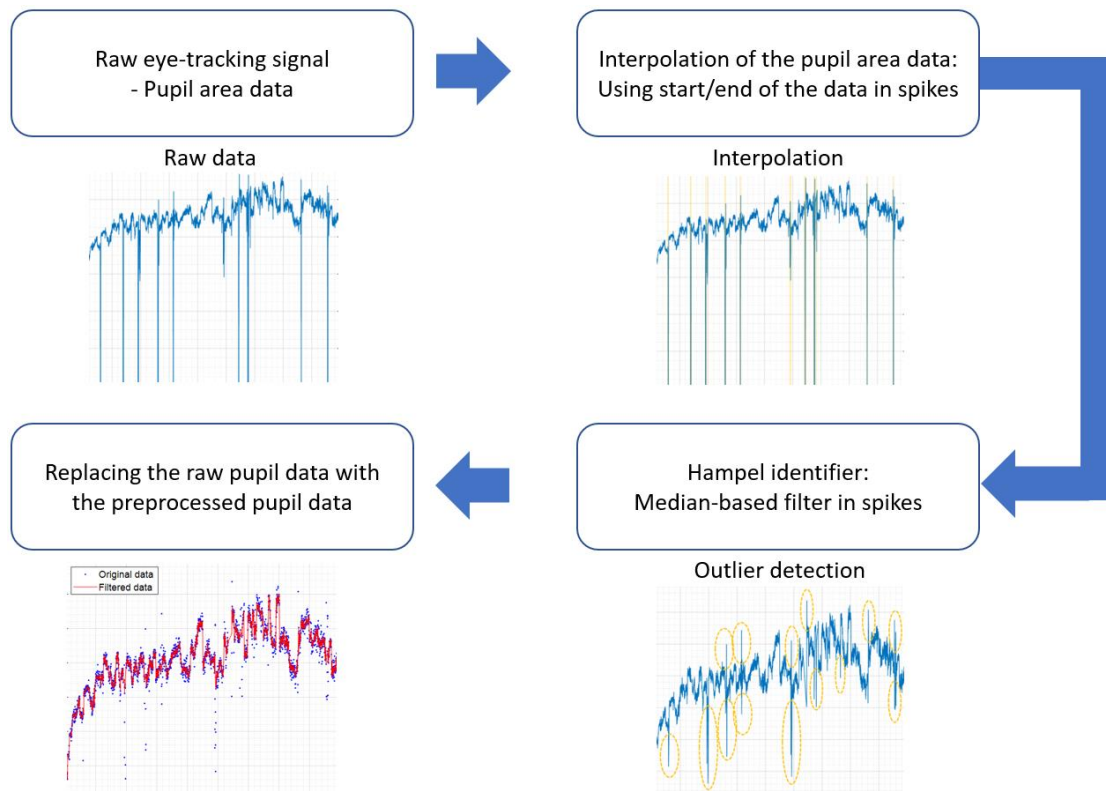


Figure 11. Raw pupil area data (Left) and processed interpolation of pupil area data in saccades (Right)

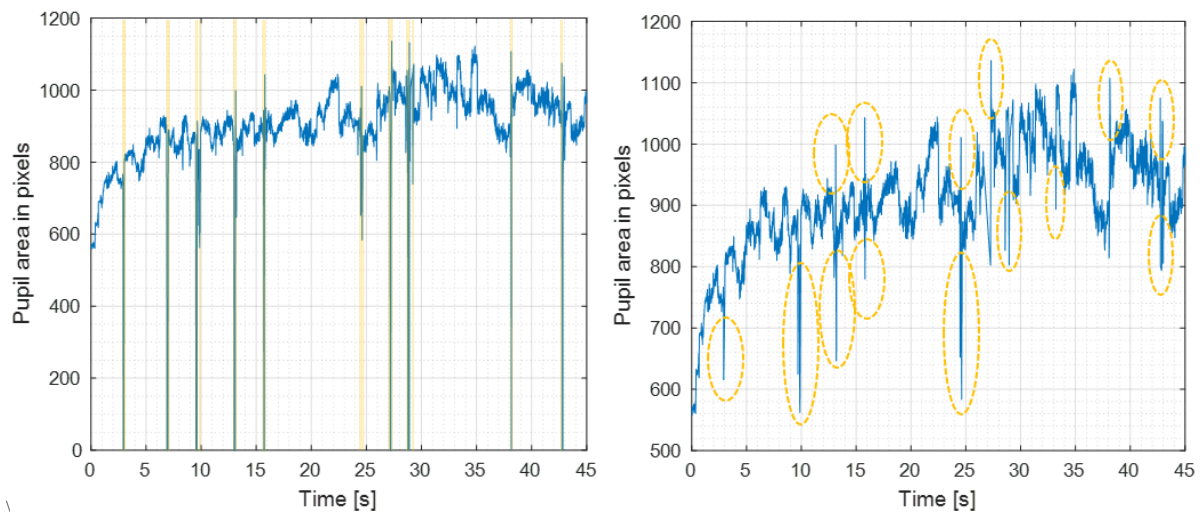
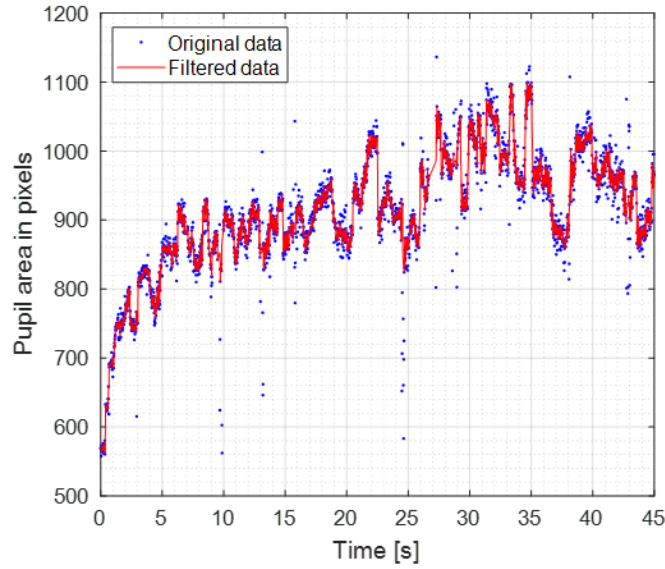


Figure 12. Hampel identifier: outlier removal



Velocity and Dispersion Threshold

In the velocity threshold, a saccade is classified when the velocity of the eye movement exceeds a certain threshold. It is around 30~40 degrees per second. Fixations are classified when the velocity is below this threshold over a certain duration > 100 milliseconds. For dispersion threshold, saccades are detected when gaze dispersion exceeds a threshold of around 1~2 degrees of visual angle, and fixations are identified when gazes stay within this dispersion threshold over 100 milliseconds. These two thresholds can be adjusted based on the specific requirements of the study or the characteristics of the subject being observed. In other research, the velocity threshold of saccadic eye movements is set over 300 deg/sec, and that of fixations is set below 100 deg/sec [188, 189]. Another research used point-to-point velocities for a velocity threshold of 20 deg/sec for sorting saccades and fixations [189]. For the dispersion threshold, another research sets the visual angle change with a minimum duration threshold of 100-200 milliseconds [190, 191].

4.3 Artifact Removal of Other Physiological Signals

In this study, other physiological signals were measured by the wrist wearable sensor, E4 Empatica. The wearable sensor can collect BVP, EDA, and skin temperature, as well as an accelerometer and a gyroscope. Figure 13 shows a brief specification of the wearable sensor. In the following sections, these physiological signals are introduced.

Figure 13. Empatica E4 specification [192]



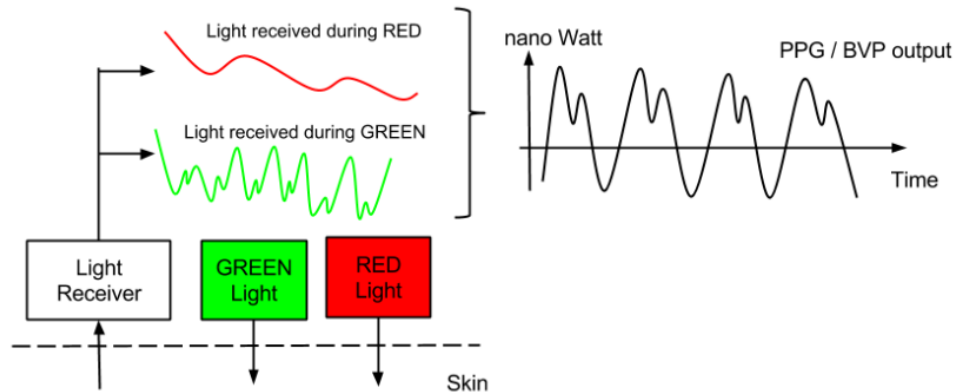
4.3.1 Blood Volume Pulse (BVP)

Blood Volume Pulse (BVP), also known as Photoplethysmography (PPG), is a non-invasive method used to measure blood flow dynamics and heart rate. Figure 14 illustrates the basic principle of how PPG is measured [193]. In the PPG measurement setup, Green and Red LEDs emit light towards the skin. These lights penetrate the skin and are absorbed differently by the blood depending on the wavelength.

- The Green LED light is absorbed more by the blood, which reflects the pulsatile component of blood volume and contains most of the pulse wave information, e.g., heartbeats.
- The Red LED light penetrates deeper and is less influenced by blood volume changes, reflecting the baseline (slow changes) in blood volume as a reference light level to cancel motion artifacts.

A portion of the emitted light is then reflected from the skin and underlying blood vessels and is detected by a light receiver. The reflected light signals vary with each heartbeat as blood volume changes dynamically with the cardiac cycle. The light receiver captures these variations in reflected light intensity and converts them into an electrical signal which is producing the PPG waveform. This waveform as shown in the figure, represents the changes in blood volume over time. The waveform shows periodic peaks corresponding to each heartbeat, where the amplitude and frequency of the signal provide information about the cardiovascular system, such as heart rate and blood flow patterns. This data can be used to compute other feature data such as Inter-Beat Interval (IBI) and Heart Rate (HR).

Figure 14. Detection theory of PPG sensor and its output [193]



For computing IBI and HR, PPG waveforms are properly processed/analyzed as shown in Figure 15. The diastolic points that are the local minima in the PPG waveform represent the lowest blood volume within a cardiac cycle and are essential for calculating IBI. Systolic points, the local maxima, indicate the peak blood volume following ventricular contraction and can be used to estimate vasoconstriction, offering insights into vascular tone and conditions like hypertension. The dicrotic notch, a small downward deflection following the systolic peak, represents the brief rise in aortic pressure as the aortic valve closes and is particularly useful for studying different cardiac diseases, especially in static subjects. Following this, the dicrotic wave, a secondary upward wave, results from the dicrotic notch and reflects arterial elasticity and wave reflection phenomena in the circulatory system. However, these key points can be distorted and corrupted by motion artifacts and may become undetectable. P1 to P3 in Figure 16 are examples of points corrupted by artifacts. From the data shown, these points are not computable for IBI and HR due to the irregular and unpredictable behaviour of PPG signals. Therefore, we designed an algorithm to estimate the key points for HR and IBI as much as possible. Based on estimated regular IBI intervals before P1~P3, we could calculate the average IBI. This average IBI is used as a criterion to determine whether a diastolic or systolic point is irregular. As illustrated in the example, in regions where irregular key points are found, there are multiple local minima and maxima that could be potential irregular points. Under typical conditions, such as in the absence of arrhythmias or ectopic beats, significant changes in IBI can occur within milliseconds, especially during physical activities, stress or relaxation. For example, during deep breathing or stress responses, changes in IBI can exceed 100 milliseconds between consecutive beats but less than 200 milliseconds [194]. Based on this approach, we can evaluate the IBIs obtained from potential diastolic points P1, P2 and P3. Since our experiments exclude physical exertion and do not involve participants with pathological conditions, we allowed for IBI differences of less than 200 milliseconds as previously described. For example, in the evaluation from dp1 to pd2, if each of the IBI values p12 and p23 differ from the previous average IBI by more than 200 milliseconds, these IBIs would be classified as

artifacts and therefore excluded from the analysis. However, if p13 in the IBI that is calculated from the start of P12 to the end of P23 falls within the normal variance of IBI (< 200 milliseconds), then p13 can be measured as a valid IBI along with dp1 and pd2. This algorithm enabled us to measure HR and IBI in real time. If the IBI in a given time interval does not meet these criteria, it is not recorded.

Figure 15. The PPG waveform with annotated key points: Diastolic point (local minima), Systolic point (local maxima), Dicrotic notch and Dicrotic wave. These points provide valuable information for assessing cardiovascular function including the estimation of heart rate, vascular resistance and potential cardiac abnormalities (Original image from [195]).

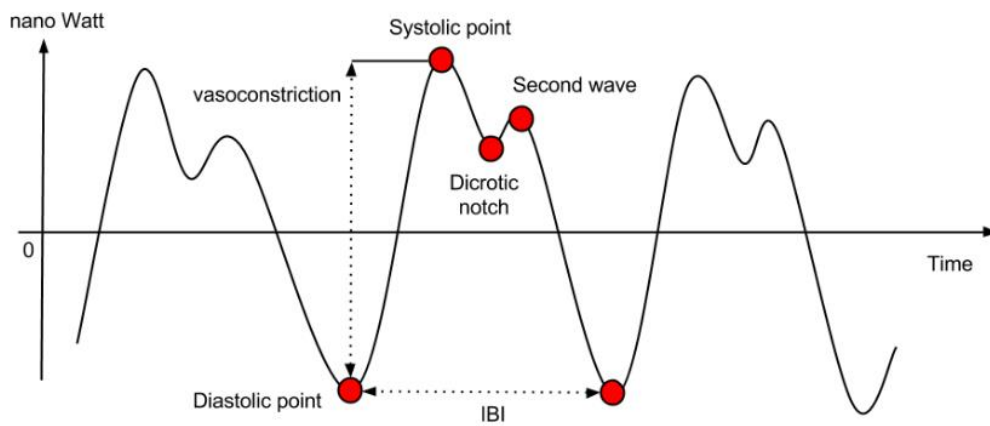
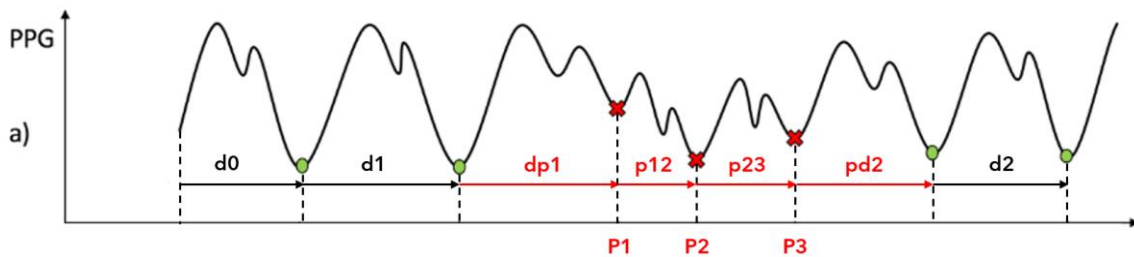


Figure 16. The PPG waveform with regular and irregular diastolic points

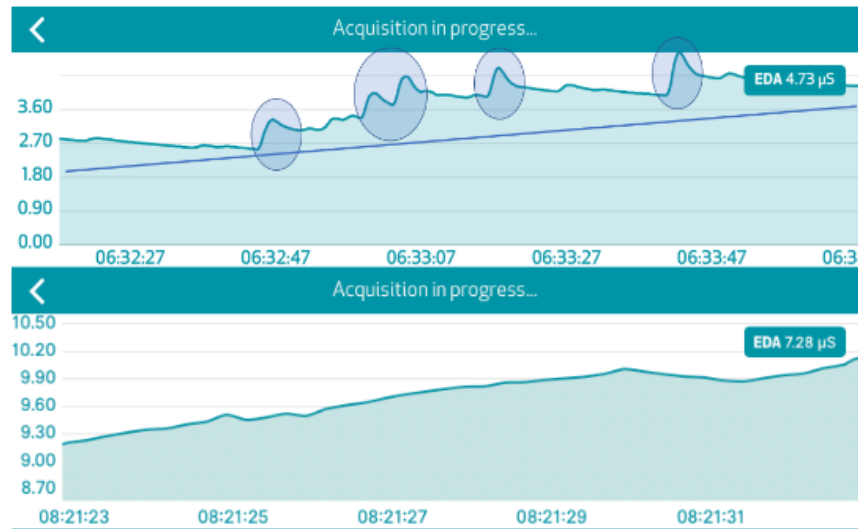


4.3.2 Electrodermal Activity (EDA)

Electrodermal activity (EDA) refers to the changes in electrical properties of the skin that occur when it receives signals from the brain via the autonomic nervous system. In response to emotional arousal, cognitive stress or physical exertion, the brain sends signals that stimulate the eccrine sweat glands, leading to an increase in sweating. Although this increase in sweat production may not always be perceptible on the skin's surface, it results in a measurable rise in skin conductance as the sweat pores begin to fill beneath

the surface. Our EDA sensor in the Empatica E4 measures skin conductance by passing a small current between two electrodes placed on the skin. This method allows for the capture of electrical conductance that is the inverse of resistance and is expressed in microSiemens (μS). The sensor is designed to detect subtle changes in skin conductance that can be indicative of physiological responses to emotional or cognitive stimuli. Skin conductance measurements are traditionally categorized into two types: tonic skin conductance level and phasic skin conductance response. Phasic skin conductance measurements are associated with short-term, discrete environmental stimuli such as sensory inputs or cognitive processes like anticipation and decision-making. These measurements manifest as abrupt increases in skin conductance. It typically feature a rapid rise followed by a slower decay as illustrated by prominent peaks shown in Figure 17 (circles on the top figure) [196]. In contrast, tonic skin conductance refers to the baseline level of skin conductance observed in the absence of specific external stimuli. This tonic level can vary gradually over time and is influenced by factors such as psychological state, hydration and autonomic regulation. Tonic changes occur over more extended periods ranging from tens of seconds to minutes. For our experiments, there is no physical exertion, and the sampling frequency of the EDA signal is 4 Hz. Therefore, we simply apply a smoothing filter, the Savitzky-Golay filter with parameters: window length of 9 samples (~ 2.15 seconds) and a polynomial order of 2.

Figure 17. EDA signals: phasic activations are circled (Top) and slow changes with no phasic significant activity (Bottom) [196]



4.3.3 Temperature

For temperature measurement, the E4 Empatica is equipped with an optical thermometer. This sensor utilizes optical principles to gauge temperature without making physical contact. These thermometers detect and analyze the infrared radiation emitted by an object. The radiation varies with temperature according to Planck's law of black-body radiation. This non-contact measurement capability makes optical thermometers particularly valuable in situations where traditional thermometers are impractical or unsafe, such as when moving objects or in hard-to-reach areas. For collected skin temperature data, we do not apply any filter to measured data.

4.4 Computational Complexity and Scalability

I evaluated the computational complexity of our signal processing methods. For EEG recordings, I proposed two automated algorithms: revised versions of Independent Component Analysis (ICA) and Common Average Referencing (CAR). ICA has higher computational complexity than CAR due to the iterative nature of its convergence process and the matrix operations involved in decomposing multichannel EEG signals into independent components. This suggests that the computational complexity of the proposed ICA method is greater than that of the revised CAR, as the additional steps in these revised algorithms pertain solely to processing EEG signals within fixation and saccade ranges using each method's procedure separately. When comparing the dominant terms in Big-O notation analysis of ICA-based algorithms used in other real-time applications, the computational complexity of the proposed revised ICA does not differ significantly. In the revised ICA, 18 channels were selected out of 32 available channels. The preprocessing approach comprises three stages for EEG, including fixation and saccade separation, and two stages for pupil signal processing. The code was implemented on an Intel Core i7-10700 CPU (2.9 GHz) with 16 GB RAM. In the first stage of EEG preprocessing, a band-pass filter was applied to retain signals between 0.1 and 40 Hz. Using Big-O notation, the computational complexity of this stage is $O(\log_2 n)$. The second stage involves automated ICA with additional steps, including variance ratio calculations between saccade and fixation EEG signals and reconstruction of selected components, with a complexity of $O(km_{channel}^2n + m_{epoch}n + m^2n)$, where $m_{channel}$ and m_{epoch} represent the number of channels and epochs, respectively, n is the number of samples, and k is the number of ICA iterations. In the final stage, features were extracted from two selected channels (Cz and FCz) among the reconstructed signals. Each channel was segmented into four frequency ranges: Delta (0.1~4 Hz), Theta (4~7.5 Hz), Alpha (7.5~12 Hz), and Low (0.1~10 Hz), using band-pass filters, followed by feature extraction for each range. For eye-tracking signals, preprocessing utilized a Hampel filter with four steps. The first step, involving a sliding window

for median and median absolute deviation calculations, has the computational complexity of $O(nk \log_2 k)$, where k represents the sliding window size. However, since each data segment contains 500 samples and the sliding window size is significantly smaller than the samples, n , the complexity simplifies to $O(n)$. Additionally, the preprocessing of EEG and eye-tracking signals required information on fixation and saccade intervals, classified using velocity and acceleration calculations of saccadic movements with a complexity of $O(n)$. Thus, the total computational complexity for generating features to feed into the trained model is $O(km_{channel}^2n + m_{epoch}n + m^2n + n)$, where the ICA algorithm dominates ($km_{channel}^2n \gg (m_{epoch} + m^2 + 1)n$). When comparing ICA-based preprocessing algorithms in other real-time EEG applications, the Big-O notations are as follows: $O(\max(m^3n, n^2, km^2n))$ with the final complexity of $O(n^2)$ [197], and $O(\max(km^2n, m^3, m^2n))$ with the final complexity of $O(km^2n)$ [198, 199], where m and k represent the number of channels and the number of ICA iterations, respectively. The dominant term in the Big-O notation reflects how computational complexity scales with dataset size, highlighting challenges with larger samples or channels. Comparing these with our proposed algorithm's complexity, $O(km_{channel}^2n)$, the number of EEG channels, m , significantly impacts the resulting Big-O notation. Therefore, reducing the number of EEG channels or utilizing simpler algorithms than ICA-based methods could strengthen the implications for real-time applications. To ensure scalability, strategies like dimensionality reduction, limiting iterations, parallel processing, and incremental learning can balance efficiency and accuracy, making the approach practical for real-world, large-scale applications. For model training, the LightGBM, $O(nd + bd)$, is preferable to the XGBoost, $O(nd \log n)$, where n is the number of samples, d is the number of features, and b is the number of bins for LightGBM due to its relatively lower complexity and faster runtime, making it more suitable for large datasets [200].

4.5 Summary

The various signals introduced above are recorded to collect participants' physiological responses across multiple experiments. Artifacts in the collected signals are cleaned through the preprocessing algorithms previously described to correct and minimize artifacts. In particular, the integration of EEG and eye-tracking data enables us to improve preprocessing algorithm performance for artifact removal, e.g., keeping salient signal sources better while removing artifacts effectively. Therefore, the implementation of a more sophisticated artifact removal algorithm utilizing both modalities highlights the potential of this integration. In this chapter, research questions 1~2 are fully addressed by suggesting advanced artifact removal algorithms and comparing them with conventional algorithms for EEG and eye-tracking signals. Then, the preprocessed data is used to extract various features representing participants' physiological responses. The

responses are subsequently utilized to develop AI models that assess physiological states and cognitive decision-making processes such as error detection, risk detection, and user preferences.

4.6 Limitations

Although our signal-processing approach demonstrates promising computational efficiency, its generalizability across different datasets and conditions requires further validation. To enhance its potential for real-time applications, the preprocessing steps should be tested with fewer EEG channels while maintaining effective artifact removal. Additionally, future studies should validate the approach using varied datasets, including different sampling rates, etc, and explore adaptations for lower-resource devices and domain adaptation techniques to address variability across diverse datasets. This could further enhance our approach's scalability, applicability and generalizability, making it more versatile in real-world contexts.

Chapter 5. Detection of Braking Activation and Different Deceleration Degrees Related to Vehicle Speed at Collision

In this chapter, our primary goal is to develop user perception models regarding proximity to a potential collision subject and other behaviours, such as alerting, lane change for collision avoidance using data features from an eye-tracking system and various physiological recording systems such as EEG, BVP, GSR, and skin temperature for detecting different potential collision subject types and classifying varied conditions that can be perceived as different severity levels by participants. Additionally, we include other events such as alerting, full stop, and avoidance of collisions. By using CARLA, an advanced open-source simulation software for autonomous driving research, we create various scenarios in expressway and small-town areas that are integrated with factors such as vehicle speed, number of vehicles/pedestrians, and road settings. In these designed traffic scenarios, we collect eye-tracking data and physiological signals to extract features of each scenario. Using the extracted features, we developed classification models to detect multiple braking activation timings and deceleration rates based on human perception of proximity between a driving vehicle and a potential collision subject, such as vehicles and pedestrians. These factors can influence human perception of potential risk severity and are utilized to design various traffic scenarios representing multiple classes in the user perception model. Additionally, other conditions, such as alerting and avoidance by lane changes, are considered. However, when it comes to risk, individual thresholds for what constitutes a risky situation may vary, as perceptions of safe braking timing and deceleration rates differ significantly among individuals. To address this, this research explores the use of eye-tracking and physiological signals to build user perception models under varied potential collision scenarios, adopting a personalized model approach. This approach mitigates the impact of individual differences in risk perception by ensuring that the physiological signals and features extracted from these signals are not influenced by other individuals' data. The differences in human physiological responses and extracted features arising from varying risk perceptions of individuals minimize impact on the model built for each individual.

5.1 Experiment Design

In our pursuit to devise experimental scenarios to identify varied potential risk scenarios, we employed CARLA software, as shown in Figure 18, a state-of-the-art open-source simulation software specifically designed for autonomous driving research. It provides a comprehensive and customizable environment that enables the simulation of a wide array of driving scenarios. One of the key features of CARLA is its ability to simulate both urban and rural settings, thereby providing a realistic and diverse range of environments

for testing. It also allows for the simulation of various weather conditions, enabling researchers to test how autonomous vehicles perform under different atmospheric conditions. Moreover, CARLA is equipped with a flexible API that allows users to control all aspects related to the simulation, e.g., traffic generation, pedestrian behaviours, weather conditions and much more. It also supports the configuration of diverse sensor suites, e.g., LIDARs, multiple cameras, depth sensors and GPS. In essence, CARLA serves as a powerful tool in our research by providing us with a highly detailed and adaptable platform for simulating various driving scenarios.

Figure 18. CARLA simulation environment



5.1.1 Braking Time and Distance

To determine the braking distance and total stopping time for a vehicle travelling at a certain velocity, e.g., 100 km/h, we need to consider both the reaction time and the braking distance. The reaction time is the period it takes for the driver to perceive a hazard and begin a braking action, while the braking distance is the distance the vehicle travels after the brakes are applied until it comes to a complete stop. The braking reaction time ranges from 0.36 to 2.55 seconds, and its average is around 1.175 seconds depending on the complexity of situations, distraction, etc [201, 202]. The reaction and braking distance can be calculated using the equation provided by the American Association of State Highway and Transportation Officials (AASHTO) as follows [201]:

$$\text{Braking distance [m]} = \frac{v^2}{254 \times (f + G)}$$

$$\text{Reaction distance [m]} = 0.278 \times t \times v$$

where t is the reaction time in seconds, v is the speed of the vehicle in km/h, f is the coefficient of friction, and G is the slope of the road, typically assumed to be 0 for a flat road.

The friction coefficient is an important indicator of braking effectiveness influenced by various factors such as road conditions, tire quality, and vehicle speed. Anti-lock Braking Systems (ABS) and Electronic Stability Control (ESC) can optimize the friction coefficient significantly, resulting in improved braking performance. These systems have become a de facto standard in the automotive industry due to factors such as consumer demand, market dynamics, voluntary agreements between organizations, safety ratings and requirements for Advanced Driver Assistance Systems (ADAS). All of these factors indirectly encourage the integration of advanced braking systems in commercial vehicles. Moreover, as a regulatory mandate, the National Highway Traffic Safety Administration (NHTSA) requires ABS on all new commercial vehicles with a gross vehicle weight rating (GVWR) exceeding 10,000 pounds [203]. Braking optimization systems like ESC and ABS maximize the friction coefficient to stop a vehicle as quickly as possible, often achieving a friction coefficient of up to 0.9 [204]. For instance, a vehicle travelling at 90 km/h with a friction coefficient of 0.9 requires approximately 35 meters to come to a complete stop. The required braking time can be calculated based on a deceleration rate determined by various factors. Furthermore, during emergency braking, a peak friction coefficient of up to 1.02 was measured at 40 mph, and the ESC system was enabled during road testing [203]. Assuming a friction coefficient of around 0.9, this corresponds to a deceleration rate of approximately 8.83, and the braking time can be calculated using the equation below.

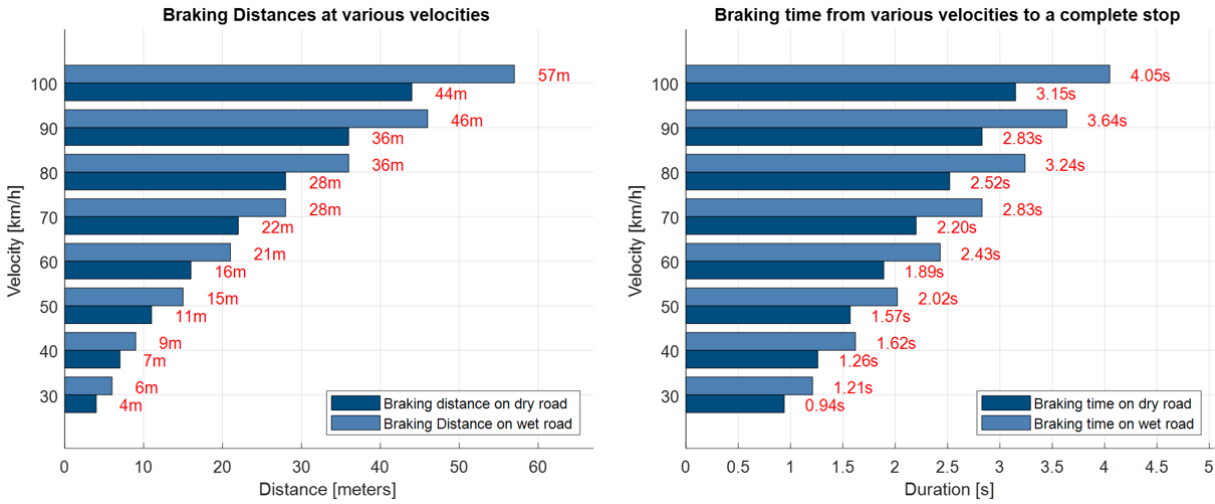
$$\text{Deceleration rate } [m/s^2] = -9.81 \times f$$

$$\text{Braking time } [s] = (v_f - v_i) / \text{Deceleration rate}$$

where f is the friction coefficient, v_f is the final velocity of the vehicle, assuming 0 as a complete stop, v_i is the initial speed at a brake start.

For fully braking from 90 km/h, it takes approximately 2.83 seconds on dry road conditions. Figure 19 shows the required distances and durations to stop vehicles at various speed conditions on dry road conditions (However, it depends on various factors such as other road conditions, vehicle conditions including weights, brake conditions, etc.). In this study, we designed scenarios excluding wet road conditions to minimize the effects of varied friction coefficients. However, there is a driver reaction time that requires a total of distance calculations till a fully stop condition.

Figure 19. Braking distance and time at various vehicle speeds



5.1.2 Design of Traffic Scenarios

For participants, all traffic scenarios do not involve actual driving at any point during the experiment. Instead, the experiment focuses on using self-driving features, with scenarios designed to include cases of potential collision risk. Each participant evaluates whether the self-driving vehicle activates the braking system at an appropriate timing, considering the proximity between the driving vehicle and the potential collision subject ahead. As previously mentioned, participants can perceive potential risks up to 6~8 seconds in advance. To maintain attention and focus on the situation, visual alerting is provided before this perception window as part of the experimental design. In line with this, related references highlight that the design standards of the American Association of State Highway and Transportation Officials (AASHTO) allocate 1.5 seconds for perception time and 1.0 second for reaction time at most, but normal reaction-duration time is between 0.75 and 1.5 seconds depending on depending on the alertness, fatigue level, alcohol consumption, and age of the driver [205, 206]. The 1.5-second perception duration is sufficient for physiological signals, such as EEG and pupil responses, to exhibit reactions to visual stimuli. These include responses such as P1, N170, P200, P300, and N400, which are analyzed at intervals of 100 ms [207, 208]. Such EEG analyses provide evidence of the visual information processing occurring in the brain. Moreover, the recognition of pedestrians and vehicles in front is sufficiently possible for participants under highway conditions approximately 6–8 seconds prior to a potential collision. Visual alerting regarding approaching potential risk subjects is provided even earlier, ensuring that participants have ample time for perception and decision-making regarding the proximity and speed of the vehicle. As a vehicle approaches a potential collision, drivers typically perceive an increasing need to activate the braking system. This perception is influenced by factors such as the time-to-collision (TTC) and the relative speed between the driver's vehicle

and the obstacle. Studies have shown that drivers' reactions and perceptions are significantly affected by the speed of an approaching obstacle, with higher speeds leading to faster recognition of the necessity to brake [209, 210]. Based on these factors, the experiment ensures that participants have sufficient time and cues to recognize potential risks effectively. This setup implies that participants can be well-equipped to evaluate potential risks within the designed scenarios. For more details in designing experiment scenarios, there are multiple factors considered such as vehicle speed, the number of vehicles/pedestrians involved in scenarios, the number of lanes, and road type/area settings such as a small-town area and a highway. It is because traffic volumes, speed, visibility and pedestrian volume also significantly affect likelihood and consequence of pedestrian crashes [211]. These factors enable us to simulate various conditions. Table 12 lists these factors for each scenario. Figure 20 shows some of captured scenes that represent the classes regarding potential collision and non-collision risks in user perception models. In the road settings, there are highway and small town conditions, with the highway condition further divided into two road environments: the first lane and the exit lane. Each road condition has a maximum vehicle speed, and scenarios are designed to include potential collisions with pedestrians or vehicles while using self-driving features. Additionally, participants may react differently to potential collision risks in each scenario based on the number or type of pedestrians involved. However, due to relative vulnerability, scenarios involving a child (as opposed to an adult) as the collision subject, multiple pedestrians rather than a single one, or collisions occurring at higher speeds rather than lower speeds may result in increased fatalities [211, 212, 213, 214]. These considerations imply that both relative and absolute perspectives of risk were taken into account when designing scenarios in this study, which helps participants better perceive and assess risk. For example, in the context of pedestrians and collisions, other researchers have highlighted that relative risk provides a more accurate picture of the risk to individual pedestrians, whereas absolute risk gives a clearer understanding of the total number of lives lost and injuries incurred, irrespective of usage. Focusing on only one type of risk may lead to a misinterpretation of true risk [211, 215].

Table 12. Three traffic environments with various factors

Factors	Environment 1	Environment 2	Environment 3
Road settings	Highway	Highway	Small town
Driving lane #	First lane	Exit lane	One lane
Collision subject	Pedestrian or Vehicle	Vehicle	Pedestrian
Type of subject	Adult/Kid	Sedan	Adult/Kid
Number of pedestrians	Single / Multi (5)	Vehicle	Single / Multi (5)
Max. vehicle speed	84~86 km/h	84~86 km/h	30~35 km/h

Event types	Alerting / Stop / Avoidance / Collision
Deceleration rate	Three deceleration rates for severity levels

Figure 20. Captured scenes from experiments.

(a) Three road environments from left to right: Environment 1 to 3 (three different lanes)



(b) Subject type: Human and vehicles (kid, adult, vehicles on road, vehicles on road shoulder)



(c) Alerting, Stop, Avoidance, and Collision starting from top left, right, bottom left and right.



For user perception models related to braking timings and deceleration rates under potential collision risks, we designed experimental scenarios with multiple classes, as outlined in Tables 13.1 to 13.3. Various factors, introduced in Table 12, were used to define these multi-class scenarios. As introduced earlier, collision severity can vary based on pedestrian types, the number of pedestrians, and vehicle speeds [211, 212, 213, 214, 215]. In the traffic scenarios designed for this study, multiple cases are used as labels in the predictive models. For pedestrian collision cases, there are two pedestrian types, two pedestrian numbers, and two deceleration rates, resulting in a total of eight cases (derived from a $2 \times 2 \times 2$ combination). For other scenarios, there are full stop and avoidance cases, which include various combinations of pedestrian types and numbers, as well as alerting cases. Some specific examples include a collision involving one adult pedestrian at maximum vehicle speed, a full stop case, and a collision involving decelerated speed. Additionally, scenarios include multiple adults instead of one, or a child pedestrian instead of an adult, with corresponding collision cases. In the collision scenarios, physiological signals are extracted starting 3.5 seconds before the threshold of the safe braking point. Participants can recognize the collision subject up to approximately 10 seconds in advance and can perceive the timing and degree of deceleration within about 1 second before the threshold of the safe braking point. The full stop case demonstrates deceleration rates comparable to those in the two collision cases: one involving a collision at maximum speed with no deceleration (rate = 0) and another involving a collision where the vehicle decelerates by approximately 40% of the maximum speed. Thus, participants are able to perceive or anticipate upcoming potential situations based on the distance between the collision subject and the driving vehicle before the epoch extraction duration, as well as on braking system activation and deceleration rates during the epoch extraction duration. As mentioned earlier, a 1-second epoch provides sufficient time for perceiving visual stimuli and for observing physiological responses. Notably, factors such as deceleration rates can also be recognized as dynamic stimuli, and participants are capable of perceiving these factors within a short time frame. Furthermore, EEG responses provide evidence of participants' perceptions of such stimuli [216]. Furthermore, Amandeep et al. (2024) conducted a study to investigate how participants make utilitarian decisions (i.e., decisions that minimize overall harm) under time pressure conditions of 1–3 seconds [217]. The experiment involved a foggy road scenario where participants were required to decide the direction of avoidance when suddenly encountering pedestrians. These decisions involved avoiding collisions with either a large group of pedestrians or an elderly person. The study found that when the time pressure for making decisions regarding the suddenly appearing visual stimulus (pedestrians) was relaxed from 1 second to 2 seconds, the proportion of utilitarian decisions increased from 70% to approximately 93%. This result indicated that 2 seconds was sufficient for participants to make utilitarian decisions in this experimental setting. In my research, participants recognized a visual alerting for a potential collision subject before participants perceive the subjects in advance around 6~8 seconds at least. This provides enough time to

perceive and evaluate the risk of potential collisions as proximity between the driving vehicle and the subjects gets closer. This allows participants to be aware of the collision likelihood increases. Combined with the earlier-mentioned EEG and pupil responses, which occur within 100–300 ms post-stimulus, this demonstrates that participants can exhibit perception-related reactions within a short time frame. Other studies have reported that drivers require approximately 1.5 seconds to 2 seconds [205, 206] to make decisions, further supporting the appropriateness of the design used in this thesis. Moreover, under focused conditions, humans can perceive changes in visual stimuli even more rapidly [218, 219]. This reinforces the idea that the experimental design provides sufficient time for participants to recognize and respond to potential collision risks effectively.

Table 13.1 Defined classes from pedestrian-related situations in Traffic Environment 1

Pedestrian(s) at risk			
Risk	Severity 0	Severity 1	Severity 2
Pedestrian(s)	Full stop	50~60 km/h	Around 85 km/h
1 adult	a1_sv_0	a1_sv_1	a1_sv_2
5 adults	a5_sv_0	a5_sv_1	a5_sv_2
1 kid	k1_sv_0	k1_sv_1	k1_sv_2
5 kids	k5_sv_0	k5_sv_1	k5_sv_2
Vehicle(s) at risk			
Risk	Severity 0	Severity 1	Severity 2
Vehicles	Full stop	50~60 km/h	Around 85 km/h
Vehicle on road	collision_sv_0	collision_sv_1	collision_sv_2
Vehicle on shoulder	parked_side_sv_0	parked_side_sv_1	parked_side_sv_2
No risk			
No risk types	Alerting, ‘Beeping’	Driving w/o risk	N/A

- ‘parked_side’: vehicles parked on the shoulder

- ‘collision_’: Vehicle collision

Table 13.2 Defined classes from pedestrian-related situations in Traffic environment 2

Collision with vehicle(s) at risk				
Risk	Severity 0	Severity 1	Severity 2	Severity 3
Vehicles	Full stop	< 30 km/h	50~60 km/h	Over 90 km/h
Vehicles on road	collision_sv_0	collision_sv_1	collision_sv_2	collision_sv_3

No risk			
No risk types	Alerting, ‘Beeping’	Driving w/o risk	Collision avoidance, ‘AV’

- Collision avoidance ‘AV’: No collision and lane change

Table 13.3 Defined classes from pedestrian-related situations in Traffic environment 3

Pedestrian(s) at risk			
Risk	Severity 0	Severity 1	Severity 2
Pedestrian(s)	Full stop	10~15 km/h	Around 30 km/h
1 adult	a1_sv_0	a1_sv_1	a1_sv_2
5 adults	a5_sv_0	a5_sv_1	a5_sv_2
1 kid	k1_sv_0	k1_sv_1	k1_sv_2
5 kids	k5_sv_0	k5_sv_1	k5_sv_2
Vehicle(s) at risk			
Risk types	Dangerous driving 1	Dangerous driving 2	N/A
No risk			
No risk	Alerting, ‘Beeping’	Driving w/o risk	Pedestrian (sidewalk)

5.1.3 Participants

Ten healthy adults with normal or corrected-to-normal vision (7 men and 3 women) residing in Waterloo and Toronto (Ontario, Canada) were recruited for this experiment. The participants' ages ranged from 28 to 34 years, with an average age of 30.8 (SD $\pm$ 2.6). All participants had no history of ocular or neurological diseases, disorders, or surgeries. Before participation, each participant provided informed consent either through written documentation or verbally in accordance with the protocols approved by the Office of Research Ethics at the University of Waterloo (Application No. 43926). The existing studies indicated that a strong response of pupil dilation and ERN/ErrPs was observed, effectively detecting differences between experimental conditions and increasing proportionally with the intensity of unexpected stimuli, such as surprise [10, 16, 30]. These findings suggest that the experimental design and responses are capable of producing a large effect size. Based on power analysis, a large effect size ($d \geq 0.8$) represents a substantial and meaningful difference or relationship that can be detected with fewer participants [220, 221]. For within-subject designs, a large effect size ($d \geq 0.8$) requires only 10 participants, which aligns with the conditions of our experiment. Furthermore, other existing studies on error detection using EEG signals recruited only 7 to 10 participants [11, 12, 148]. However, to improve the generalizability of the model, it is essential to increase the number of participants, which remains one of the limitations of this study.

5.1.4 Experiment Procedure

As previously described, a detailed procedure was designed for this experiment. Before starting an experimental session, participants were equipped with eye-tracking systems, EEG systems and wrist-wearable sensors. Once the devices were in place, their signals were checked and calibrated to ensure accurate recording from each modality. Subsequently, participants were asked to drive a vehicle for approximately 5-10 minutes to familiarize themselves with the appropriate braking timings before the session officially began. Each participant completed 9 sessions, each consisting of approximately 20-30 trials. After each session, participants were provided with ample break time to rest as needed. However, breaks were limited to no more than 10 minutes to prevent the overall experiment duration from becoming excessively long. At the end of each session, the data recording system was reviewed to ensure consistent signal quality. If necessary, the eye-tracking system was recalibrated to maintain accuracy.

5.2 Data Analysis

First, the simplest model would be one that can perform binary classification to determine whether a potential risk exists or not. For instance, regardless of vehicle speed, the number of subjects involved or the severity of potential risks, this model would be simply built to categorize situations as either risky or non-risky. This straightforward approach allows the model to quickly and efficiently assess potential dangers in various driving scenarios.

5.2.1 Signal Preprocessing and Feature Extraction

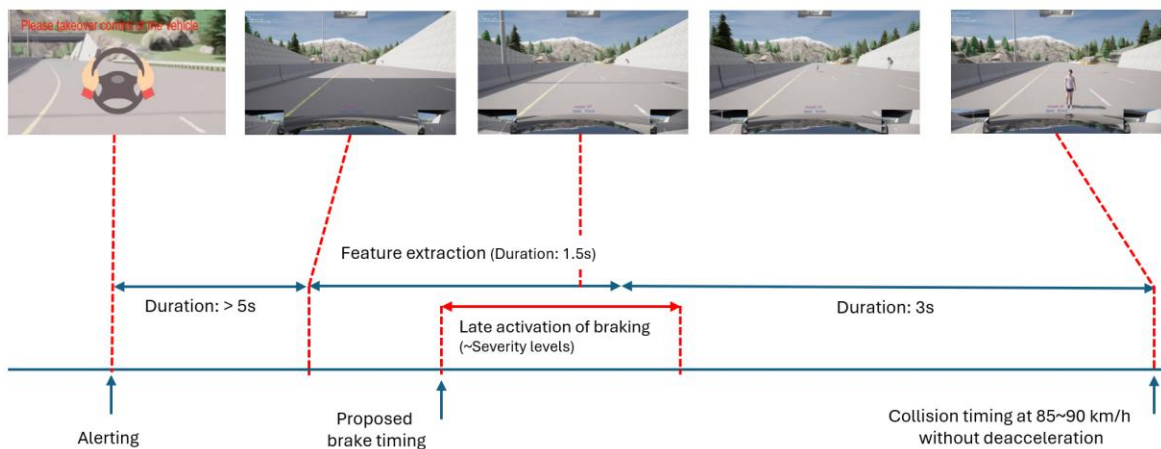
To remove artifacts of collected physiological signals and eye-tracking data, we used the same approach as Chapter 4. This preprocessing process is selected across raw signals in this research. Also, we detail the preprocessing algorithms along with the types of features extracted and the methods employed for feature extraction.

5.2.2 Feature Epoch

In this study, it is vital to determine when participants begin to perceive collision subjects, proximity between the driving vehicle and potential collision subject and changes in vehicle speeds, e.g., braking timing and deceleration rates, which have an impact on individual perception of risk severity. Identifying the precise moment when participants recognize the change and the proximity while driving towards risky points is challenging and can vary among individuals. To effectively extract epochs for feature analysis, we employed a two-step approach. The first step defines the time ranges for epoch extraction based on critical

driving events. The second step involves feature extraction focused on gaze fixation data within these specified ranges. Given the maximum driving speed in our experiment, approximately 84-86 km/h, a full stop requires around 3.2-3.5 seconds. As illustrated in Figure 21, we initiated the braking process approximately 3.5 seconds before reaching a designated collision point by considering a maximum driving speed in traffic scenarios. To gather physiological signals relevant to feature extraction, we set the recording duration from 4.5 seconds before the anticipated full stop to 3.5 seconds before reaching the collision point under full stop conditions. In scenarios involving collisions, we delayed the physiological signal recording slightly beyond the proposed braking initiation. This delay allows participants to perceive the delayed activation of the braking system and a clearer understanding of the emerging risk. This chosen duration captures epochs where risk perception is more acute as the vehicle approaches the collision point. Unlike non-deceleration conditions where deceleration has not yet begun, the onset of braking in this context heightens the perception of changes in vehicle speed, which can be associated with the perception of risk severity. Features are extracted using recorded physiological signals within a 1-second window based on gaze fixation data. Adjusting the epoch start timing according to fixation data maximizes the relevance and accuracy of the analysis. In addition to determining the epoch start timing, selecting an appropriate epoch duration is equally critical. This is because the choice of epoch duration for feature extraction can directly affect the quality of the extracted features and the performance of machine learning models. In this study, the epoch duration plays a key role in ensuring that physiological responses to stimuli, such as those related to braking events or collision risks, are accurately captured and aligned with key moments, such as the safe braking threshold.

Figure 21. Traffic scenario design/procedure



While longer durations generally ensure reliable capture of physiological responses, shorter durations introduce limitations and require a minimum threshold for accurate data extraction. As noted in Section

2.2.3, EEG and eye-tracking signals require a minimum duration of 200–300 ms, whereas heart activity studies suggest a minimum of 900 ms. To accommodate heart activity features, this study adopted an epoch duration of 1000 ms. However, excluding heart activity features could allow for significantly shorter durations, optimizing the system for real-time performance. This consideration is particularly relevant when interpreting performance metrics across diverse traffic environments and classification scenarios, as the ability to adjust epoch duration to specific conditions could further enhance the adaptability and accuracy of classification models. In the previous chapter, ‘EEG artifact removal,’ we compared the characteristics of EEG data obtained during gaze fixation with those measured during random timings with the same fixation duration after preprocessing to minimize artifacts. The results demonstrated more stable features in the EEG data acquired during gaze fixation. Based on this finding, we adopted an epoch setting in this experiment that relies on gaze fixation.

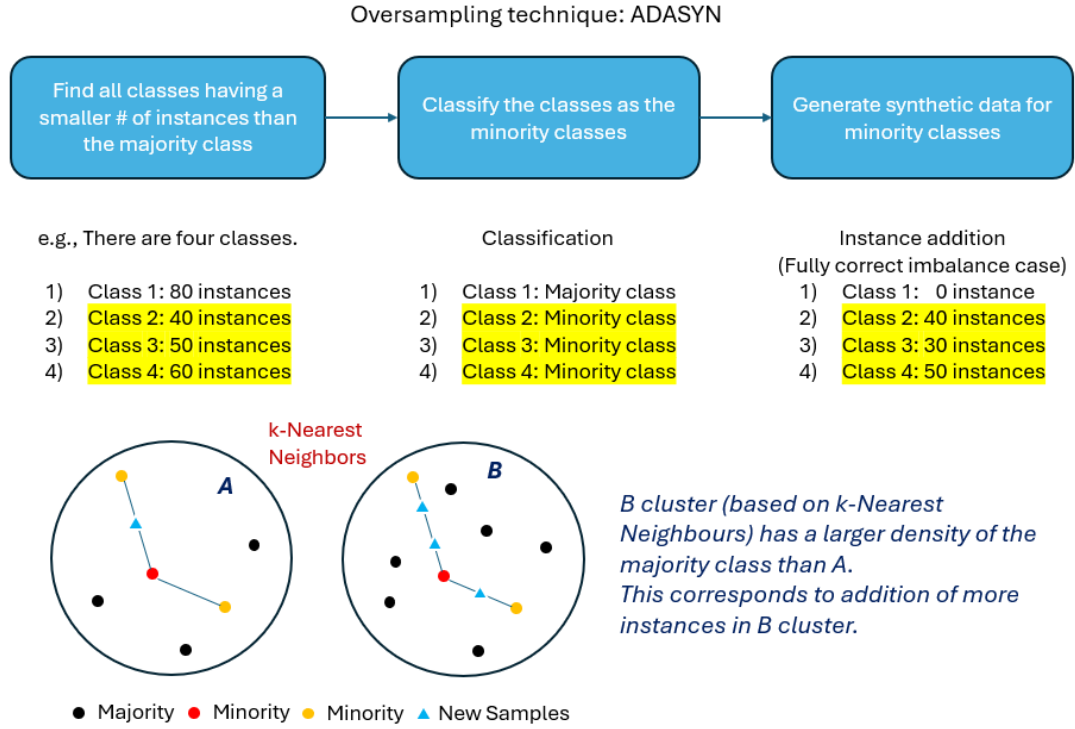
5.2.3 Class Imbalance

For this study, various classes introduced in Table 13.1~3 are used for ML model training. During data processing, we notice that there are a few cases having 10~50% fewer instances than the majority class, e.g., 2 participants have fewer instances for ‘parked_side_sv_1’ and ‘parked_side_sv_2’ in Table 13.1. We first balance the number of instances for each class to match the majority class. Next, we randomly downsample the instances to the average number across all classes before oversampling. This step helps mitigate the impact of oversampled instances on the ML model training. For oversampling, we utilized the Adaptive Synthetic Sampling (ADASYN) technique. ADASYN handles imbalanced datasets by creating synthetic instances for the minority class. Unlike the Synthetic Minority Over-sampling Technique (SMOTE), which creates synthetic instances uniformly across the minority class, ADASYN focuses on hard-to-classify instances near the decision boundary between the minority and majority classes. Therefore, it can add more instances in the hard-to-classify areas and reduce overfitting risks. However, ADASYN can be sensitive to noise as it may create instances in ambiguous regions.

5.2.4 Feature Analysis Methods

To identify the most relevant features in our dataset, we employed a variety of feature importance analysis methods. These methods help in understanding the significance of different features by evaluating their impact on the predictive power. The selected methods ensure a comprehensive evaluation. The following subsections describe each method in detail.

Figure 22 illustrates a procedure for balancing class numbers and how ADASYN works.



Analysis of Variance (ANOVA)

ANOVA is a statistical method used to compare the means of three or more samples. In the context of feature importance, it assesses the variance between groups and within groups to determine whether the means of different groups are significantly different. It is commonly used in the analysis of continuous variables. ANOVA is based on the F-test that is calculated as:

$$F = \frac{MS_{\text{between}}}{MS_{\text{within}}}$$

where MS_{between} (Mean Square Between) and MS_{within} (Mean Square Within) are calculated by $\frac{SS_{\text{between}}}{df_{\text{between}}}$ and $\frac{SS_{\text{within}}}{df_{\text{within}}}$.

- SS_{between} (Sum of Squares Between) = $\sum_{i=1}^k n_i (\bar{X}_i - \bar{X})^2$ and
- SS_{within} (Sum of Squares Within): $\sum_{i=1}^k \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2$
- $df_{\text{between}} = k - 1$ (degrees of freedom between groups)
- $df_{\text{within}} = N - k$ (degrees of freedom within groups)

Chi-squared Test

The Chi-squared test is used to determine if there is a significant association between categorical variables. For feature importance, it measures the independence of two variables by comparing observed and expected frequencies. The Chi-squared statistic is calculated as:

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

where O_i is the observed frequency and E_i is the expected frequency.

If the data contains continuous features, discretize them through binning. Next, construct a contingency table that displays the frequency of each combination of feature and target categories. Then, calculate the Chi-squared statistic (χ^2) for each feature to measure its association with the target variable. Based on these values, compute the p -values to evaluate the statistical significance of the association, where a lower p -value (e.g., $p < 0.05$) indicates a stronger relationship. Finally, rank the features by their χ^2 scores or p -values, selecting the most significant features for further analysis or model training. This process helps identify the features with the strongest predictive power in a simple and statistically robust manner.

Mutual Information

Mutual Information measures the amount of information obtained from one variable through another. In feature selection, it quantifies the dependency between features and the target variable. Mutual Information between two discrete random variables X and Y is defined as:

$$I(X; Y) = \sum_{x \in X} \sum_{y \in Y} p(x, y) \log \left(\frac{p(x, y)}{p(x)p(y)} \right)$$

where $p(x, y)$ is the joint probability distribution, $p(x)$ and $p(y)$ are the marginal probability distributions.

BorutaSHAP

BorutaSHAP is a feature selection method that combines the Boruta algorithm with SHAP (SHapley Additive exPlanations). It iteratively removes features to find the ones that are important for prediction. Based on cooperative game theory, the SHAP value is calculated as:

$$\phi_i(v) = \sum_{S \in N \setminus \{i\}} \frac{|S|!(n - |S| - 1)!}{n!} (v(S \cup \{i\}) - v(S))$$

where S is a subset of all features N and $v(S)$ is the value of the model trained with feature subset S .

In the Boruta algorithm, an ML model algorithm is used to assess the feature importance by comparing original and shadow features created by shuffling original features. The default ML algorithm is random forests in the Boruta algorithm.

Least Absolute Shrinkage and Selection Operator (LASSO)

Lasso is a regression analysis method that performs both variable selection and regularization to enhance prediction accuracy. It adds a penalty equal to the absolute value of the magnitude of coefficients. The Lasso optimization problem is:

$$\min_{\beta} \left(\frac{1}{2N} \sum_{i=1}^N (y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij})^2 + \lambda \sum_{j=1}^p |\beta_j| \right)$$

where y_i is the target variable, x_{ij} are the features, β_j are the coefficients, and λ is the regularization parameter.

Permutation Importance

Permutation importance measures the importance of a feature by evaluating the increase in prediction error after randomly permuting the feature's values. This breaks the relationship between the feature and the target, allowing the assessment of the feature's contribution. For a given feature X_j , the permutation importance is:

$$PI(X_j) = \frac{1}{K} \sum_{k=1}^K \left(\text{Loss}(\hat{f}, X_j^{\text{perm}}, X_{\text{train}}^j) - \text{Loss}(\hat{f}, X_j, X_{\text{train}}) \right)$$

where $\hat{f}$ is the trained model and X_j^{perm} is the permuted feature.

By applying these diverse methods, we were able to comprehensively evaluate the importance of each feature in our dataset. ANOVA and Chi-squared tests provided insights into the variance and association of features with the target variable, respectively. Mutual Information offered a measure of dependency between features and the target. BorutaSHAP combined robustness and interpretability, highlighting the most relevant features. LASSO contributed by performing both variable selection and regularization, while Permutation Importance assessed the contribution of each feature by its impact on prediction error.

Together, these methods ensured a thorough and reliable selection of features, enhancing the predictive accuracy and robustness of models.

5.2.5 Various Granularity Levels in Three Traffic Environments

In this section, we discuss the user perception classification of collision risk and non-collision cases at various granularity levels, as outlined in Tables 14.1, 14.2, and 14.3. Each table provides a detailed breakdown of different classes under various traffic environments and levels of granularity, which are crucial for developing robust user perception models encompassing various collision risk conditions. Table 14.1 presents the user perception classifications for Traffic Environment 1 across five different granularity levels. At Level 2, the classification is divided into two primary categories: Human and Vehicle. Within each category, situations are classified as either no collision (sv_0) or collision (sv_1 and sv_2). As the granularity level increases to Level 3, the classification becomes more detailed, incorporating scenarios with pedestrian and deceleration factors. At Level 4, the classification reaches its highest granularity, distinguishing between various pedestrian conditions and two deceleration rates (e.g., a1_sv_1, k1_sv_1). This comprehensive model accounts for braking actions, including timings, deceleration rates, and collision subject conditions, enabling varied perceptions of different risk severity and non-risk cases.

Table 14.1 Collision case classification of various granularity levels in Traffic Environment 1

Granularity level [the number of classes]				
Class Level	Lv 1 [2]	Lv 2 [4]	Lv 3 [8]	Lv 4 [10]
Type	Pedestrians			
No collision	No collision: 'sv_0'	No collision 'sv_0'	N/A	No collision 'sv_0'
		Alerting 'Beeping'		Alerting 'Beeping'
Collision	Collision: 'sv_1' & 'sv_2'	Collision: 'sv_1'	'a1_sv_1'	'a1_sv_1'
			'a1_sv_2'	'a1_sv_2'
			'k1_sv_1'	'k1_sv_1'
			'k1_sv_2'	'k1_sv_2'
		Collision: 'sv_2'	'a5_sv_1'	'a5_sv_1'
			'a5_sv_2'	'a5_sv_2'
			'k5_sv_1'	'k5_sv_1'
			'k5_sv_2'	'k5_sv_2'

- 'sv_1': severity level 1 / 'sv_2': severity level 2

Table 14.2 outlines the user perception classifications for Traffic Environment 2 with two granularity levels. At Level 1, the model is simply classified into two categories for vehicles: no collision (sv_0 and no_obj) and collision (sv_1 and sv_2). At Level 4, the classification becomes more detailed with specific conditions such as alerting or beeping and introducing the concept of autonomous vehicle-driven collision avoidance (AV). Collisions are further classified by severity levels (sv_1 and sv_2), which can provide an understanding of vehicle-related risks on the road.

Table 14.2 Collision case classification of various granularity levels in Traffic Environment 2

Granularity level [the number of classes]				
Class Level	Lv 1 [2]	Lv 2 [3]	Lv 4 [4]	Lv 4 [5]
Type	Vehicle			
No collision	No collision: 'sv_0' and 'no_obj'	No collision 'sv_0'		
		Alerting 'Beeping'		
	N/A	N/A	N/A	'AV'
Collision	Collision: 'sv_1' & 'sv_2'	Collision	'sv_1'	'sv_1'
			'sv_2'	'sv_2'

- 'sv_1': severity level 1 / 'sv_2': severity level 2

- 'Beeping': alerting potential collision risks

- 'AV': collision avoidance by autonomous vehicle driving / 'no_obj': no collision

Table 14.3 details the collision classifications for Traffic Environment 3 across three granularity levels. At Level 1, the classification focuses on humans distinguishing between no collision (sv_0) and collision (sv_1 and sv_2). Level 2 adds more details that include scenarios for alerting or beeping. At Level 3, the classification is further refined to account for different groups of humans such as adults and children with specific codes for varying severities (e.g., a1_sv_1, k1_sv_2). This detailed classification allows for a more precise prediction and understanding of human-related risks in different traffic situations.

Table 14.3 Collision case classification of various granularity levels in Traffic Environment 3

Granularity level [the number of classes]				
Class Level	Lv 1 [2]	Lv 2 [4]	Lv 3 [8]	Lv 4 [10]
Type	Pedestrians			
No collision	No collision: 'sv_0'	No collision 'sv_0'	N/A	No collision 'sv_0'
		Alerting 'Beeping'		Alerting 'Beeping'

Collision	Collision: 'sv_1' & 'sv_2'	Collision: 'sv_1'	'a1_sv_1'	'a1_sv_1'
			'a1_sv_2'	'a1_sv_2'
			'k1_sv_1'	'k1_sv_1'
			'k1_sv_2'	'k1_sv_2'
		Collision: 'sv_2'	'a5_sv_1'	'a5_sv_1'
			'a5_sv_2'	'a5_sv_2'
			'k5_sv_1'	'k5_sv_1'
			'k5_sv_2'	'k5_sv_2'

- 'sv_1': severity level 1 / 'sv_2': severity level 2

5.3 Building User Perception Models for Braking Activation/Deceleration and Other Non-Braking Actions

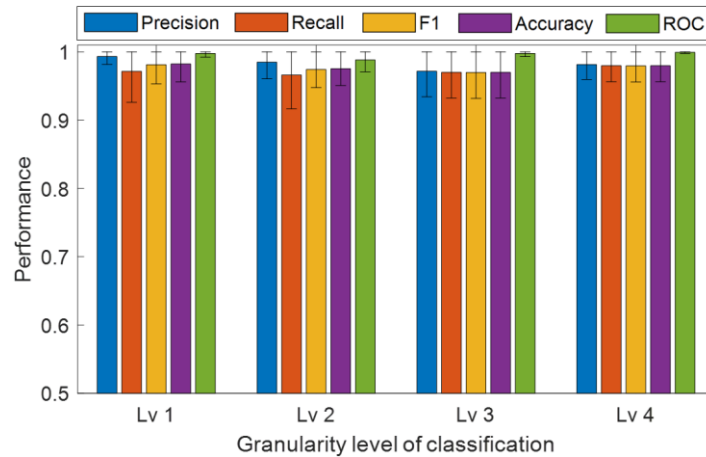
In this section, we introduce classification models with their performances built based on various conditions shown in previous sections, e.g., traffic environments, vehicle speed, braking activation timings and, deceleration rates, etc. Among the built models, the simplest model can perform binary classification to determine whether a given driving situation, either upcoming or current, is risky or non-risky. For instance, regardless of vehicle speed, potential risks ahead, or the number of subjects involved, this model would classify situations simply as risky or non-risky. Beyond binary classification, there are various classes representing each condition introduced in the previous section.

5.3.1 Performance of the User Perception Model (Traffic Environment 1)

Figure 23 presents the performance metrics for varied conditions across different granularity levels in Traffic Environment 1, ranging from Level 1 to Level 4, excluding the simplified classification of including both collisions with pedestrian(s) and vehicle. The classification model, including both pedestrian and vehicle collision cases, showed a noticeable decline in performance, where the metrics dropped to around 0.77. This drop suggests that it may significantly increase the complexity of user perception between vehicle and pedestrian collision cases. Therefore, the performance drop highlights a challenge for the model in handling the condition where class distinctions are less clear and more overlapping, potentially due to the nature of collision behaviours regardless of collision object types, e.g., pedestrians and vehicles. As a result, I developed collision risk models separately for pedestrian and vehicle collision cases. Figure 23 presents the performance of the collision risk detection models based solely on pedestrian collision cases. At Lv 1, the performance metrics are exceptionally high, nearing 1.0, indicating near-perfect classification. This

suggests that the model can accurately distinguish between the simplest classes, potential risk or no risk, highlighting its effectiveness in straightforward binary classifications. At Lv 2, the performance metrics show slightly lower than the performance at Lv 1. This slight drop reflects the increased complexity introduced by adding more detailed classes, but the model still performs very well, maintaining high accuracy, precision, recall, F1 score, and ROC values. The model performance at Lv 3 and Lv 4 was similar to that of other lower granularity models. This implies that when the classification model focuses on collision cases with pedestrian(s) only, it performs better even if the classes are more finely granulated by collision speeds and changes in the number of pedestrians. It seems to provide more apparent distinctions between these conditions. At Lv 4, more detailed classifications than Lv 3, e.g., alerting, no collision and various types of collisions with pedestrians, can also achieve slightly higher accuracy and consistency across all performance metrics.

Figure 23. Performance across various granularity levels in Traffic Environment 1.

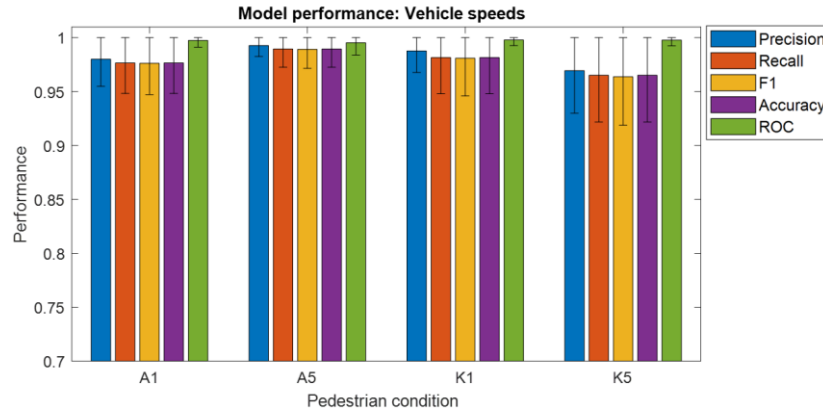


- Lv1 (2): No collision and collision.
- Lv2 (4): No collision, alerting, and collision at two different speeds.
- Lv3 (8): Collision cases (4 by 2): four pedestrian conditions (1 adult, 1 kid, 5 adults, and 5 kids) and two different vehicle speeds (med.: 45~55 km/h and high: around 85 km/h).
- Lv4 (10): No collision, alerting, and collision cases (4 by 2).

Figure 24 illustrates the model performance across four granularity classifications with vehicle speeds, depicted on the x-axis as A1 (1 adult), A5 (5 adults), K1 (1 kid), and K5 (5 kids). Each classification level consists of three classes: no collision, collision with pedestrian(s) at speed level 1 and collision with pedestrian(s) at speed level 2. The performance metrics consistently exceed 0.9 across all granularity levels, demonstrating robust classification capabilities. Performance remains stable with minor variations, suggesting the model handles different levels of granularity without significant degradation. The highest performance variability is observed in the K5 condition, which is likely due to the increased complexity of

multiple child pedestrians, yet even this condition maintains metrics above 0.9, underscoring the model's reliability. The analysis indicates that the type and number of pedestrians have a minimal impact on classification performance, and pedestrian collision conditions are highly classifiable.

Figure 24. Performance of pedestrian perception models.

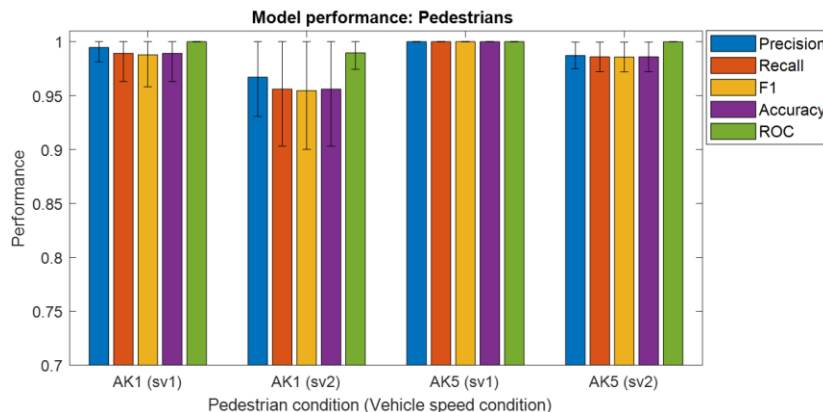


- 'A' stands for adult and 'K' stands for kid.

- The number after 'A' or 'K' stands for the number of pedestrians, e.g., 'A5' means 5 adults and 'K1' means 1 kid as pedestrian(s).

Figure 25 illustrates the model performance across various pedestrian conditions and vehicle speeds. The conditions are represented on the x-axis as AK1 (sv1), AK1 (sv2), AK5 (sv1), and AK5 (sv2) where 'A' stands for adult, 'K' stands for kid, and the number indicates the number of pedestrians. The vehicle speed conditions are denoted as sv1 (around 60 km/h) and sv2 (around 85 km/h). Each condition evaluates precision, recall, F1 score, accuracy and ROC metrics. The performance metrics consistently exceed 0.9 across all collision conditions with pedestrian(s) as the performance results in a collision with pedestrians at two vehicle speeds above.

Figure 25. Performance of pedestrian/vehicle speed perception models.

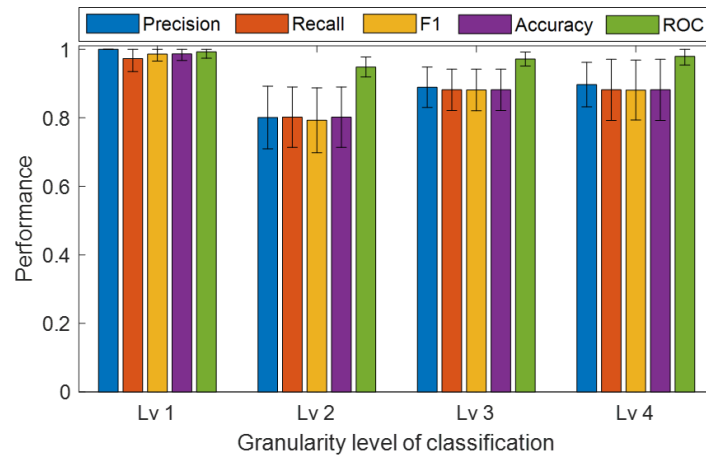


- ‘A’ stands for adult and ‘K’ stands for kid.
- The number after ‘A’ or ‘K’ stands for the number of pedestrians, e.g., ‘AK1’ means two classes: 1 adult and 1 kid, and ‘A15’ includes 1 adult and 5 adults.
- ‘sv1’ and ‘sv2’ stand for two vehicle speeds when colliding with pedestrians (sv1: ~60km/h and sv2: ~85km/h)

5.3.2 Performance of the User Perception Model (Traffic Environment 2)

Figure 26 illustrates the performance across various granularity levels (Lv 1 to Lv 4) in Traffic Environment 2. The performance metrics are exceptionally high at Lv 1, close to 1.0, indicating near-perfect classification for distinguishing between collision and no collision. At Lv 2, there is a significant drop in all metrics falling below 0.8, reflecting low performance with more complex classifications such as alerting and collision avoidance. This decline continues at Lv 3 with metrics around 0.75, highlighting further challenges in distinguishing collision cases at different vehicle speeds. Lv 4 shows a slight recovery, but the metrics remain variable. At higher levels than Lv 1, persistent difficulties in managing detailed class scenarios are observed in terms of classification performance. ROC performance remains more than 0.95 across all levels from Lv 1 to Lv 4. This indicates that despite the challenges in precision, recall, F1 score, and accuracy, the model maintains excellent discrimination ability between classes at all levels of granularity. Overall, while the model excels in more straightforward classifications, increased granularity and complexity significantly impact its performance in terms of precision, recall, F1 score, and accuracy, though its ability to discriminate between classes remains strong, as reflected by the consistently high ROC values.

Figure 26. Performance across various granularity levels in Traffic Environment 2

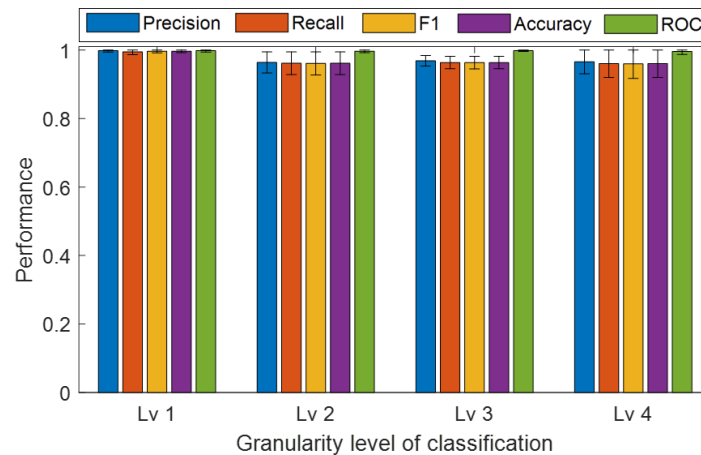


- Lv1 (2): No collision and collision
- Lv2 (3): No collision, alerting, collision avoidance by autopilot and collision
- Lv3 (4): No collision, alerting, collision cases at two different vehicle speeds
- Lv4 (5): No collision, alerting, collision avoidance by autopilot and collision cases at two different vehicle speeds

5.3.3 Performance of the User Perception Model (Traffic Environment 3)

Figure 27 illustrates the performance across various granularity levels (Lv 1 to Lv 4) in Traffic Environment 2. The metrics are exceptionally high at Lv 1, showing the perfect classification performance for distinguishing between collision and no collision. At Lv 2, these metrics slightly decline but remain robust, still above 0.9, reflecting the strong performance despite increased complexity. At Lv 3 and Lv 4, classification performance with more classes maintains similar performance levels to Lv 2, indicating no performance degradation as the number of classes (complexity) increases. ROC consistently shows an outstanding performance of over 0.95 across all levels from Lv 1 to Lv 4, indicating excellent discrimination ability between classes regardless of granularity. Overall, while the performance in precision, recall, F1 score, and accuracy slightly declines with increasing complexity from binary classification to multi-class classification, its ability to discriminate between classes remains strong, as reflected by the consistently high ROC values.

Figure 27. Performance across various granularity levels in Traffic Environment 3 (from Lv. 1 and Lv. 4)

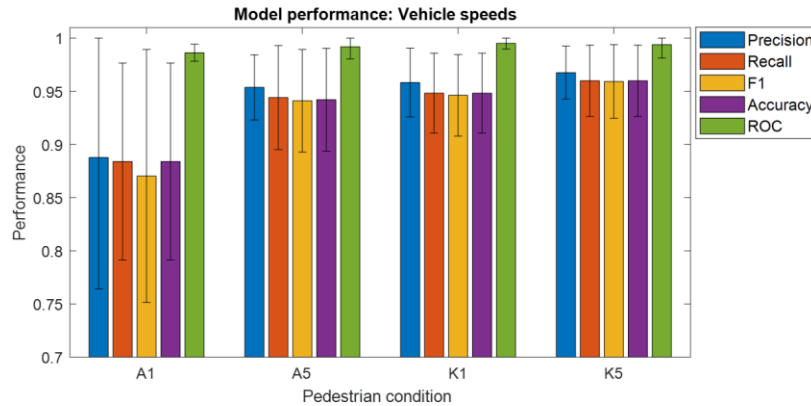


- Lv1 (2): No collision and collision.
- Lv2 (4): No collision, alerting, and collision at two different speeds.
- Lv3 (8): Collision cases (4 by 2): four pedestrian conditions (1 adult, 1 kid, 5 adults, and 5 kids) and two different vehicle speeds (extremely low: ~10 km/h and low: ~30 km/h).
- Lv4 (10): No collision, alerting, and collision cases (4 by 2).

Figure 28 illustrates the model performance for different pedestrian conditions and vehicle speeds. The pedestrian conditions are denoted as A1 (1 adult), A5 (5 adults), K1 (1 kid) and K5 (5 kids). Performance metrics consistently exceed 0.9 across most conditions, which demonstrates the robustness of the model. The highest performance is observed in the A5 condition, where all metrics are tightly grouped and close to 1.0, indicating excellent classification capabilities. For A1, while the metrics remain high, the variability is slightly more significant, reflecting some challenges in distinguishing between the two vehicle speeds

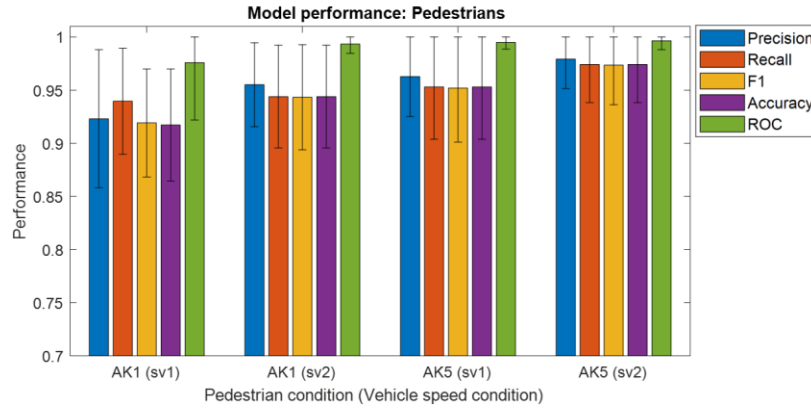
with a single adult pedestrian. In the K1 and K5 conditions, performance remains strong with all metrics above 0.9, although there is a slight decrease compared to the A5 condition. This indicates that the model performs slightly better with adult pedestrians than with child pedestrians, particularly when the number of pedestrians increases. Figure 29 illustrates the model performance for different pedestrian conditions and vehicle speeds. In Traffic Environment 3, the pedestrian conditions are denoted as AK1 (an adult and a kid) and AK5 (1 adult and 5 kids). Vehicle speeds are represented as sv1 (around 10 km/h) and sv2 (around 30 km/h). Performance metrics except for ROC consistently show high values across all pedestrian conditions, generally ranging between 0.85 and 0.95. The highest performance is observed in the AK1 (sv1) condition, where the ROC metric approaches 1.0, indicating excellent classification capabilities. For AK1 (sv2), while the metrics remain high, the variability is slightly more significant, reflecting some challenges in distinguishing between the two vehicle speeds with one adult and one child pedestrian. In the AK5 (sv1 and sv2) conditions, performance remains strong with all metrics above 0.9, although there is a slight decrease compared to the AK1 (sv1) condition. For ROC, the performance remains very high, consistently above 0.95 across all conditions, indicating excellent discrimination ability between different classes regardless of the granularity and complexity of the scenarios. This analysis suggests that while the model performs well, its ability to effectively discriminate between classes is particularly strong as reflected by the consistently high ROC values.

Figure 28. Performance of pedestrian perception models.



- 'A' stands for adult, and 'K' stands for kid.
- The number after 'A' or 'K' stands for the number of pedestrians, e.g., 'A5' means 5 adults and 'K1' means 1 kid.

Figure 29. Performance of pedestrian/vehicle speed perception models.



- ‘A’ stands for adult, and ‘K’ stands for kid.
- The number after ‘A’ or ‘K’ stands for the number of pedestrians, e.g., ‘AK1’ means two classes: 1 adult and 1 kid, and ‘A15’ includes one adult and 5 adults.
- ‘sv1’ and ‘sv2’ stand for two vehicle speeds when colliding with pedestrians (sv1: ~10km/h and sv2: ~30km/h)

Even though a finer granularity of classes increases model complexity, models with varying granularity, and therefore different levels of complexity, demonstrated high performance overall, except in cases where vehicle collision risk detection models with higher granularity performed worse than binary classification. The ability to maintain high performance despite increased model complexity suggests that models with finer class granularity may be prone to overfitting. This issue can be mitigated through techniques such as data augmentation, regularization (L1 or L2), incremental learning, reducing model complexity, and regular monitoring of model performance. Nonetheless, the overall analysis highlights the strong potential of the classification model for real-world applications in diverse pedestrian and vehicle interaction scenarios, as the models consistently maintain high performance across varying conditions. Future research could focus on refining the model to further reduce variability and enhance performance, particularly in scenarios involving larger numbers of pedestrians and higher vehicle speeds.

5.4 Feature Importance Analysis

Using various methods of feature importance analysis, we examined changes in feature importance across different factors in our experiment design, such as pedestrian types, the number of pedestrians, vehicle speeds, and types of potential risks (e.g., collision). For each factor, the results comparing the importance of features between different conditions are presented in the following sections. Through this analysis, we identified the key features and overall trends in feature importance for different conditions. This comprehensive understanding helps recognize influential features under specific conditions, thereby enhancing the robustness of our models.

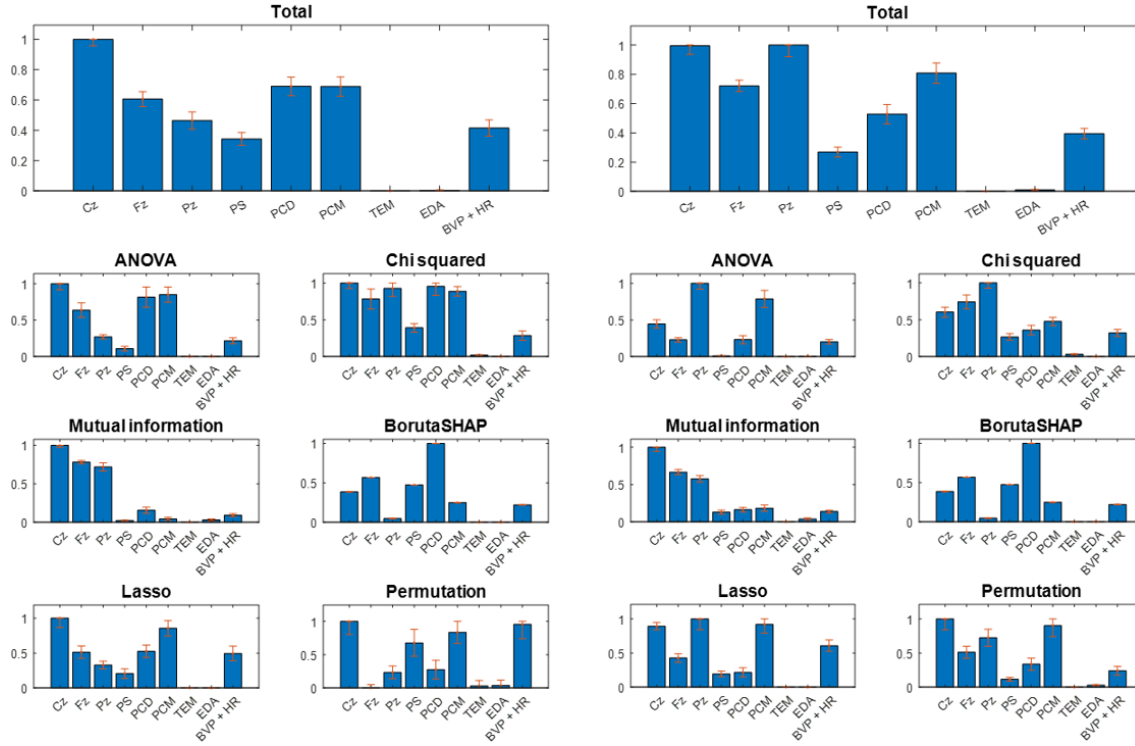
5.4.1 Feature Analysis in Scenarios: Collision with Pedestrian(s)

Figure 30 illustrates feature analysis results of the model performance for detecting two classes, ‘Adult vs Kid’ (collision with adults vs. collision with kids) and ‘Single vs Multi’ (collision with five pedestrians or five kids vs. collision with a single adult or kid), using feature importance metrics such as EEG signals (Cz, Fz, Pz), pupil size (PS), pupil center degree changes (PCD), pupil center magnitude changes (PCM), body temperature changes (TEM), electrical dermal activity changes (EDA) and heart rates (BVP + HR) in a highway condition (Traffic environment 1). In the ‘Adult vs Kid’ scenario, EEG features are highly important, particularly Cz, indicating the significance of brain activity in distinguishing between collisions involving adults and kids. Higher feature importance for the Cz channel suggests increased cognitive load and emotional response due to the perceived vulnerability difference between adult and kid pedestrians. Also, PCM and PS exhibit a high importance, indicating heightened arousal and cognitive processing for this condition. It can reflect more significant eye movement and scanning behaviour. In the ‘Single vs Multi’ scenario, features from Cz and Pz channels are the most important features with significantly higher importance in this condition, reflecting increased cognitive processing and situational awareness. PCM, which is lower than these EEG features, is also of high importance, indicating higher attentional demands that correlate with pupil-shifting magnitude changes (extensive visual scanning behaviours) while the BVP + HR feature shows similar importance levels in both conditions.

Figure 31 illustrates feature analysis results of the model performance for detecting two classes: ‘Alert vs Collision’ (alerting participants before potential risks vs. collision cases) and ‘Alert vs No collision’ (alerting participants before potential risks vs. no collision cases and full stops before pedestrian(s)/vehicle(s)). In the ‘Alert vs No collision’ scenario, the features from Cz channel are relatively important compared to other features, highlighting that the features from Cz can capture attention-related activity and decision-making, especially in tasks without dynamic interactions [222]. However, PCM is the most significant feature in this condition. Heightened vigilance by alerts may first be associated with an increased amount of visual attention/scanning correlated with the PCM feature, and the Cz features are the second important feature showing cognitive processing in detecting alerts and distinguishing them from collisions. However, PCM and PS show less importance in the ‘No Collision vs Alerting’ compared to the ‘Collision vs Alerting’ condition, but PCD feature importance shows similar in both conditions, suggesting that pupil-related features respond more strongly to ‘Alerting vs Collision’ condition. Overall, ‘No Collision vs Alerting’ shows increased feature importance in Cz, Fz, and Pz but less importance in PS and PCM. This indicates that the importance of heightened EEG features may depend on the presence of a collision as it can highly stimulate the visual cortex, even though both collision and no-collision conditions are highly associated with cognitive processing. The importance of BVP and HR is consistent between the two conditions but lower than that of EEG and pupil-related features. These results highlight the differential

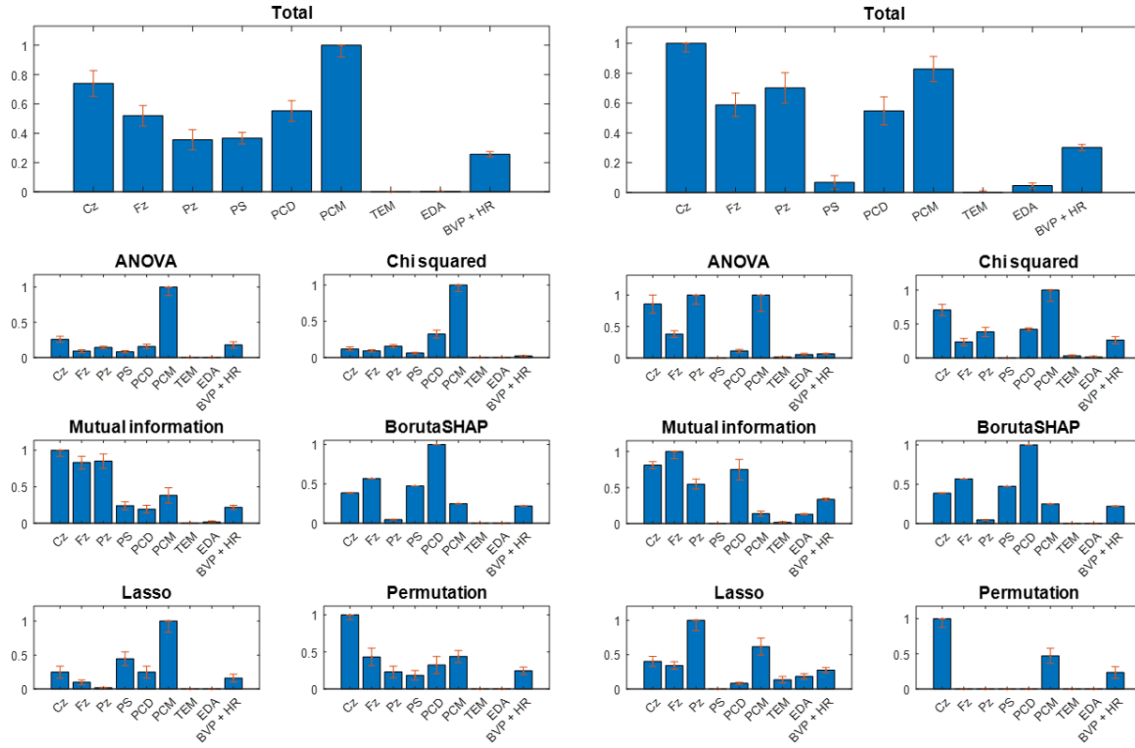
importance of EEG and pupil-related features in detecting collisions and alerting, which play a major role in both conditions.

Figure 30. Feature analysis in Traffic environment 1: pedestrian types (Left) and # of pedestrians (Right)



- (Left) Pedestrian types: adult(s) and kid(s)
- (Right) The number of pedestrians: 1 pedestrian (~Single) and 5 pedestrians (~Multi)

Figure 31. Feature analysis of three cases in Traffic environment 1: Alerting, Collision and No collision

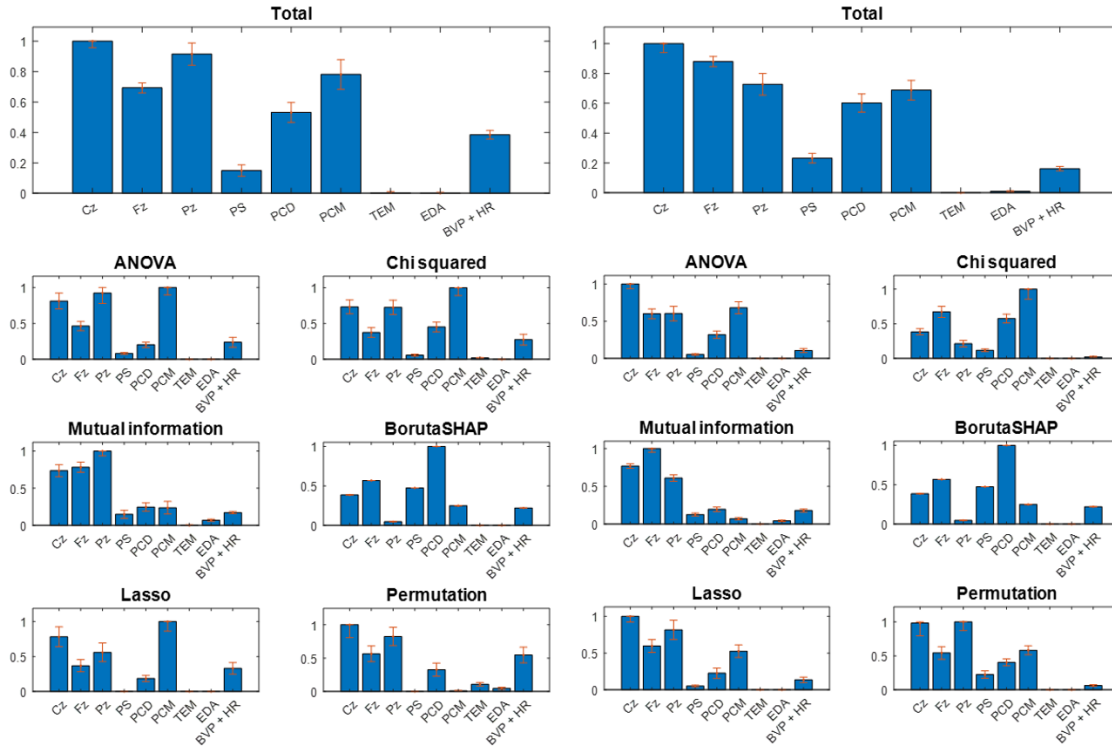


- (Left) Alerting: Alerting potential risks of collision with pedestrian(s) / Collision: Collision with pedestrian(s) and autonomous driving is fully stopped before pedestrian(s)
- (Right) Alerting: Alerting potential risks of collision with pedestrian(s) / No Collision: No collision with pedestrian(s) and autonomous driving is fully stopped before pedestrian(s)

Figure 32 illustrates feature analysis results of the model performance for detecting two classes: ‘No Collision vs Collision’ (no collision cases vs. collision cases) and ‘SV1 vs SV2’ (collision with pedestrian(s) at a speed between 50-60 km/h vs. collision with pedestrian(s) at a speed around 85 km/h). Both conditions show the same list of the top 5 features (Cz, Fz, Pz, PCM and PCD) with similar rank orders of feature importance except for the second and third-ranked features (Fz and Pz) being swapped. Cz is the most important feature in both conditions, indicating significant cognitive processing in distinguishing no-collision from collision cases and collision severity levels. It can be explained that in distinguishing visual changes, such as variations in deceleration related to risk severity levels or no deceleration, tasks under high-speed conditions may predominantly rely on sensorimotor activities (e.g., movement recognition and response). In this context, features from the Cz channel may exhibit relatively greater importance in perception compared to features extracted from other channels, such as Fz and Pz [223]. The 2nd ranked feature differs between conditions being either Fz or Pz. The frontal lobe (Fz) is involved in decision-making, motor control and visual attention, while the parietal lobe (Pz) is crucial for spatial awareness and processing visual stimuli. Research suggests that responses to visual stimuli are more pronounced in the

occipital and parietal lobes, making Pz more critical for distinguishing collision presence in conditions such as ‘No Collision vs Collision’.

Figure 32. Feature analysis of three factors in Traffic environment 1: No collision, Collision with severity



- (Left) No Collision: No collision with pedestrian(s) and autonomous driving is fully stopped before pedestrian(s) / Collision: Collision with pedestrian(s).

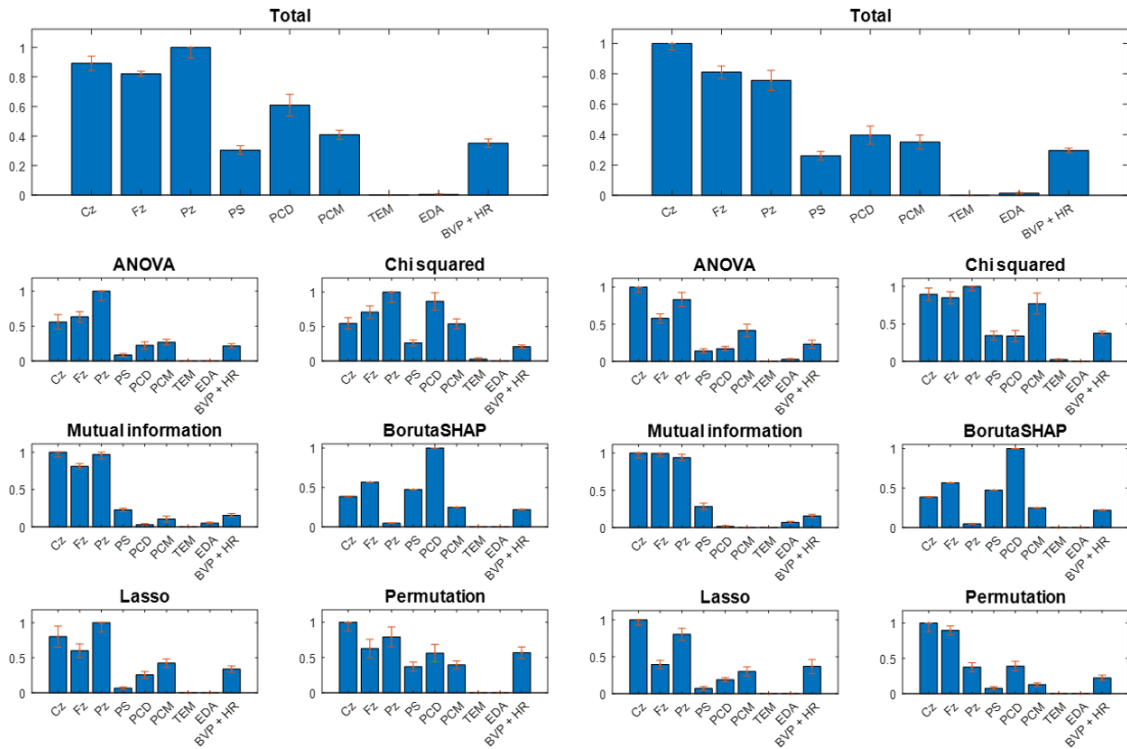
- (Right) Severity level 1~ SV1: Collision with pedestrian(s) at vehicle speed, 50~60 km/h / Severity level 2~ SV2: Collision with pedestrian(s) at a higher vehicle speed around 85 km/h than SV1.

Conversely, Fz is more significant for distinguishing between different collision speeds, particularly in high-speed and deceleration conditions, as it is closely associated with the predictive and planning functions of the frontal lobe. For pupil-related features, PCD and PCM show higher importance than PS in both conditions. However, the difference between PCM and PCD is more significant in the ‘No Collision vs Collision’ condition, possibly due to increased visual attention and scanning behaviour in the condition. In the ‘SV1 vs SV2’ condition, the feature importance of PS slightly increases in response to speed changes while that of PCM slightly decreases. TEM, EDA, and BVP + HR are less important than other features. However, BVP + HR shows higher importance than PS in the ‘No Collision vs Collision’ condition. It may reflect increased stress and arousal in collision risks, possibly more distinguishable in cardiovascular responses. Overall, this feature importance analysis is based on comparisons across different conditions. However, in cases where the model's complexity significantly increases, various factors may influence the results of the current feature importance analysis.

Traffic Environment 3

The analysis of feature importance across various conditions in Traffic Environment 3, which represents traffic situations in a small town with lower vehicle speeds (around 10-15 km/h for SV1 and around 30 km/h for SV2), reveals some interesting trends. For ‘Adult vs Kid,’ the total feature importance indicates that EEG channels are vital, as shown in Figure 33. In traffic environment 1, Cz features are highly prioritized compared to other channels in the figure. However, in traffic environment 3, Pz features show the highest importance. At Traffic Environment 3, pupil-related features in both pedestrian type and the number of pedestrian conditions (‘Adult vs Kid’ and ‘Single vs Multi’) show lower importance than the importance in Traffic Environment 1.

Figure 33. Feature analysis of two factors in Traffic environment 3: pedestrian types (Left) and the number of pedestrians (Right)



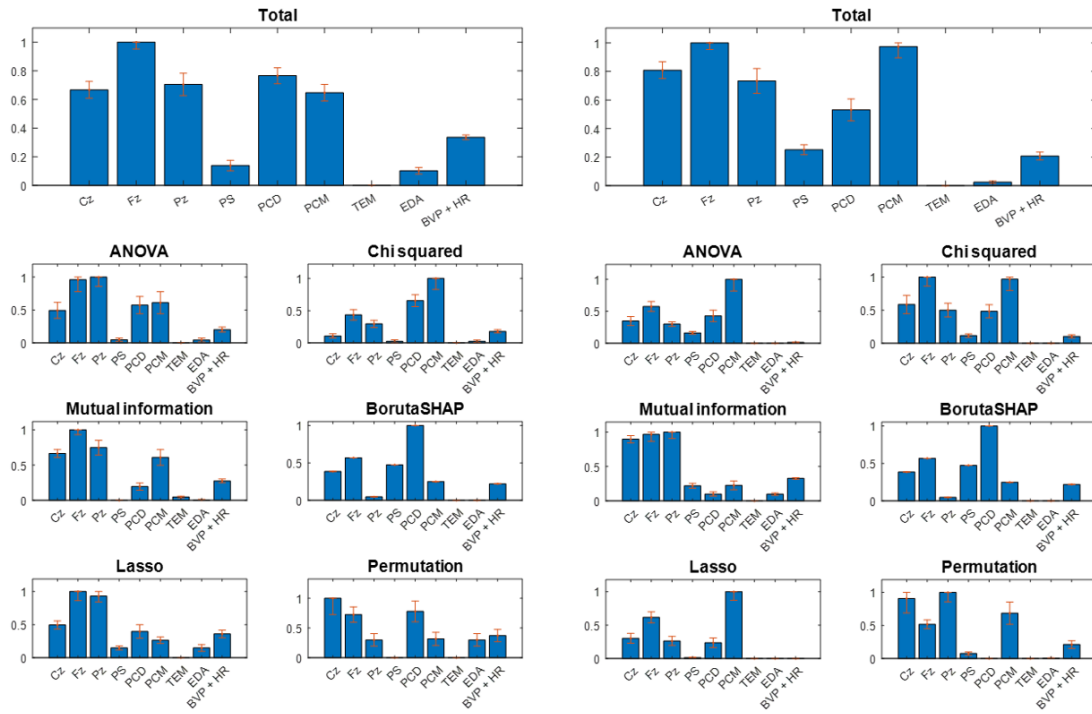
- (Left) Pedestrian types: adult(s) and kid(s)

- (Right) The number of pedestrians: 1 pedestrian (~Single) and 5 pedestrians (~Multi) regardless of pedestrian types

Figure 34 illustrates the feature importance of distinguishing between alerting participants before potential risks (Alerting) and collision cases. For both ‘Alerting vs Collision’ and ‘Alerting vs No Collision,’ the total feature importance reveals that Fz is the most significant feature, indicating that frontal lobe activity is vital for distinguishing between these conditions. This high importance of Fz suggests that the features

can show differences between attention control and visual impact in collision/potential risks, such as whether alerting is compared with the existence of collision risks. Pz and Cz follow in importance, highlighting their importance in spatial, visual information, and cognitive processing. One of the pupil-related features, PCM, highly stands out in ‘Alerting vs No Collision’ of Traffic Environment 3, but in ‘Alerting vs Collision,’ the feature importance of PCD and PCM remains similar. It may be explained by differentiated pupil dynamics in both conditions, e.g., higher correlation pupil dynamics with higher magnitudes of pupil shift in ‘No Collision’ and similar importance of PCD and PCM for distinguishing between ‘Collision’ and ‘Alerting’ cases. This result suggests that feature importance can vary depending on the traffic environment and vehicle speed. This variation is particularly notable in the ‘Alerting vs Collision’ condition. The reason is that ‘Collision’ scenarios in traffic environment 1 involve higher vehicle speeds, more dynamic changes in visual stimulation and a more significant impact on cognitive processing due to increased potential collision risks compared to traffic environment 3.

Figure 34. Feature analysis of three cases in Traffic environment 3: Alerting, Collision and No collision

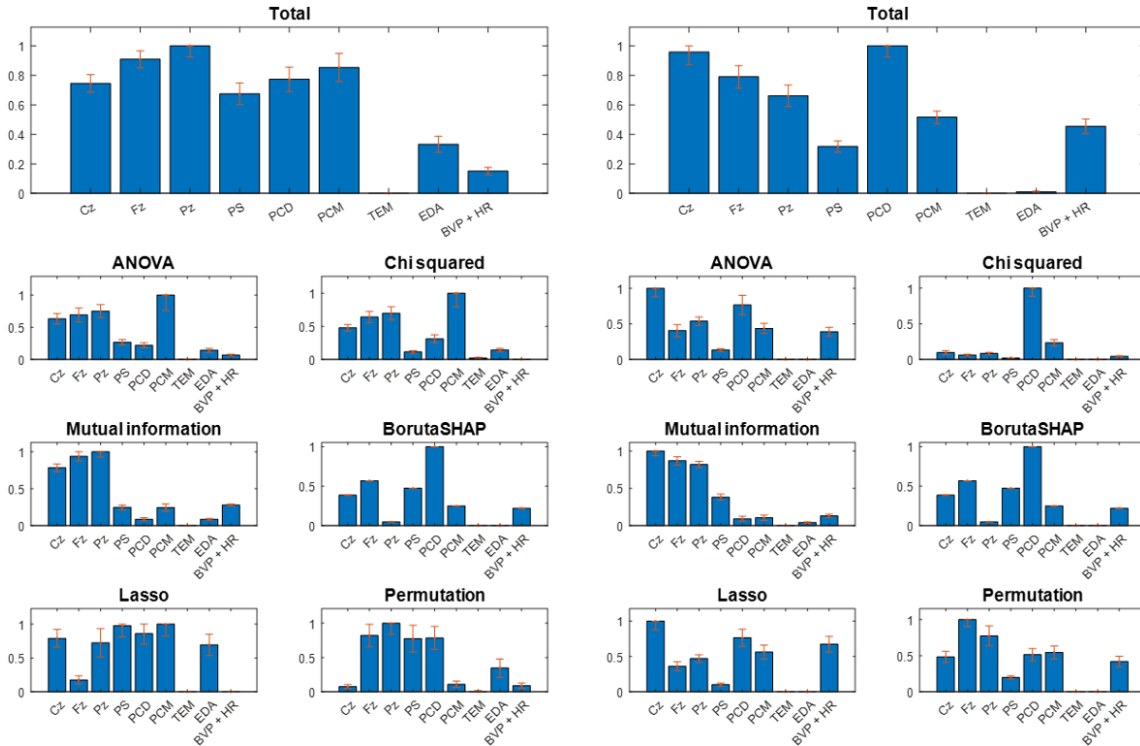


- (Left) Alerting: Alerting potential risks of collision with pedestrian(s) / Collision: Collision with pedestrian(s) and autonomous driving is fully stopped before pedestrian(s)
- (Right) Alerting: Alerting potential risks of collision with pedestrian(s) / No Collision: No collision with pedestrian(s) and autonomous driving is fully stopped before pedestrian(s)

The analysis of feature importance across various conditions in Traffic Environment 3 reveals distinct physiological and neural responses for the ‘No Collision vs. Collision’ and ‘SV1 vs. SV2’ conditions in

Figure 35. In the ‘No Collision vs. Collision’ scenario, the total feature importance indicates that EEG features and pupil-related features are essential. Overall, feature importance results are increased across EEG and pupil-related features in Traffic Environment 3 compared to Traffic Environment 1. This increase may be due to the longer exposure duration to visual targets, which means that the lower vehicle speeds can provide enough periods to evoke physiological responses. This extended duration ends up enhancing the distinctiveness of overall feature importance, resulting in heightened feature significance in the ‘No Collision vs. Collision’ condition. In the ‘SV1 vs. SV2’ scenario, the total feature importance trend is similar to that observed in Traffic Environment 1 except for the PCD and BVP + HR features. The slower collision speeds cause longer exposure to colliding objects (pedestrians) and increase the discriminative power of features related to heart activity and scanning behaviours. The EEG features are ranked after PCD, and their importance ranks are opposite to the ranks of EEG features in the ‘No Collision vs Collision’.

Figure 35. Feature analysis of three factors in Traffic Environment 3: No collision, Collision, and Collision severity.



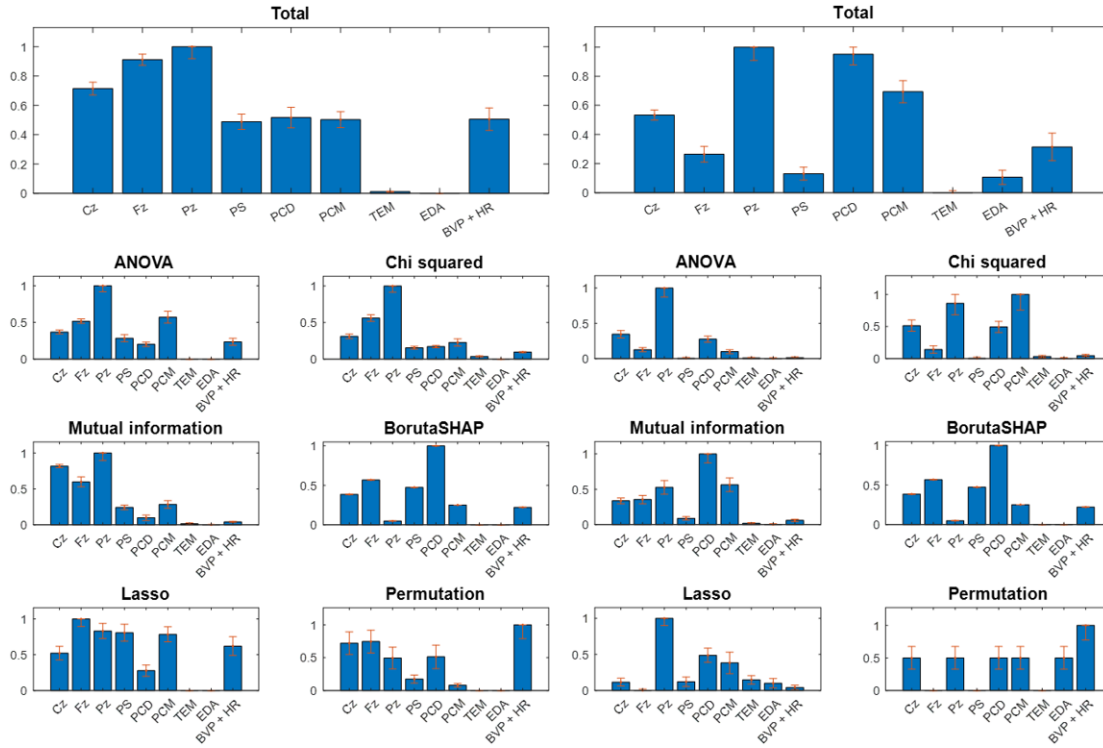
- (Left) No Collision: No collision with pedestrian(s) and autonomous driving is fully stopped before pedestrian(s) / Collision: Collision with pedestrian(s) and autonomous driving is fully stopped before pedestrian(s)
- (Right) Severity level 1~ SV1: Collision with pedestrian(s) at vehicle speed, 10~15 km/h / Severity level 2~ SV2: Collision with pedestrian(s) at vehicle speed around 30 km/h

Overall, when comparing feature importance across various pedestrian collision scenarios, the importance of features from pupil dynamics generally increased under highway conditions (high traffic speed) compared to low traffic conditions (e.g., small towns). Additionally, the difference in traffic speed (vehicle speed) appeared to modulate the feature importance of both EEG and pupil features. This observation suggests that under high-speed conditions, pupil responses may become more pronounced compared to low-speed conditions, potentially explaining the observed changes in feature importance. Existing studies may indirectly imply that there are feature importance changes and adjustments between pupil and EEG features as changes in vehicle/traffic speeds [224, 225, 226].

5.4.2 Feature analysis in scenarios: Collision with vehicles

Figure 36 illustrates the importance of distinguishing between alerting participants before potential risks (Alerting) and collision risks with other vehicles (Collision) in Traffic Environment 1. In the ‘Alerting vs Collision’ scenario, the total feature importance is significant for EEG features compared to other features. This result differs from the feature importance observed in scenarios involving collision risks with pedestrians in the same traffic environment. For instance, in vehicle collision risk conditions, all pupil features (PS, PCD and PCM) show similar importance. However, in pedestrian conditions, differences in the importance of three pupil features are shown to be more evident. In the ‘Alerting vs No Collision’ condition, PCD shows higher importance than PCM for potential collision risks with vehicles. For potential collision risks with pedestrians, PCM shows higher than PCD regardless of traffic speed conditions, which are opposed to each other. It indicates that depending on collision object types, pupil shift behaviours can be shown differently, e.g., more distinguishable eye movement in magnitude and degree. In EEG features, features from the Pz channel show the highest importance in both ‘Alerting vs Collision’ and ‘Alerting vs No Collision’ conditions. EEG feature importances do not show any trend in the collision risk conditions of pedestrians and vehicles. The average of the EEG features is higher than that of the pupil features in the ‘Alerting vs Collision’ condition, but they show similar in the ‘Alerting vs No Collision’ condition. The feature importance of BVP + HR is higher in the ‘Alerting vs Collision’ condition than in the ‘Alerting vs No Collision’ condition, and the importance is higher than in pedestrian collision risks regardless of collision and no collision. The above results highlight a distinct pattern in the relative importance of physiological and visual processing features between vehicle and pedestrian collision risks. As shown in Figure 36, the importance of features in the ‘Alerting vs No Collision’ condition is highly selective compared to ‘Alerting vs Collision.’ The top three features are Pz, PCD, and PCM in ‘Alerting vs No Collision,’ but all EEG features are the top three features in ‘Alerting vs Collision.’

Figure 36. Feature analysis of three factors in Traffic Environment 1: Alerting, No Collision and Collision

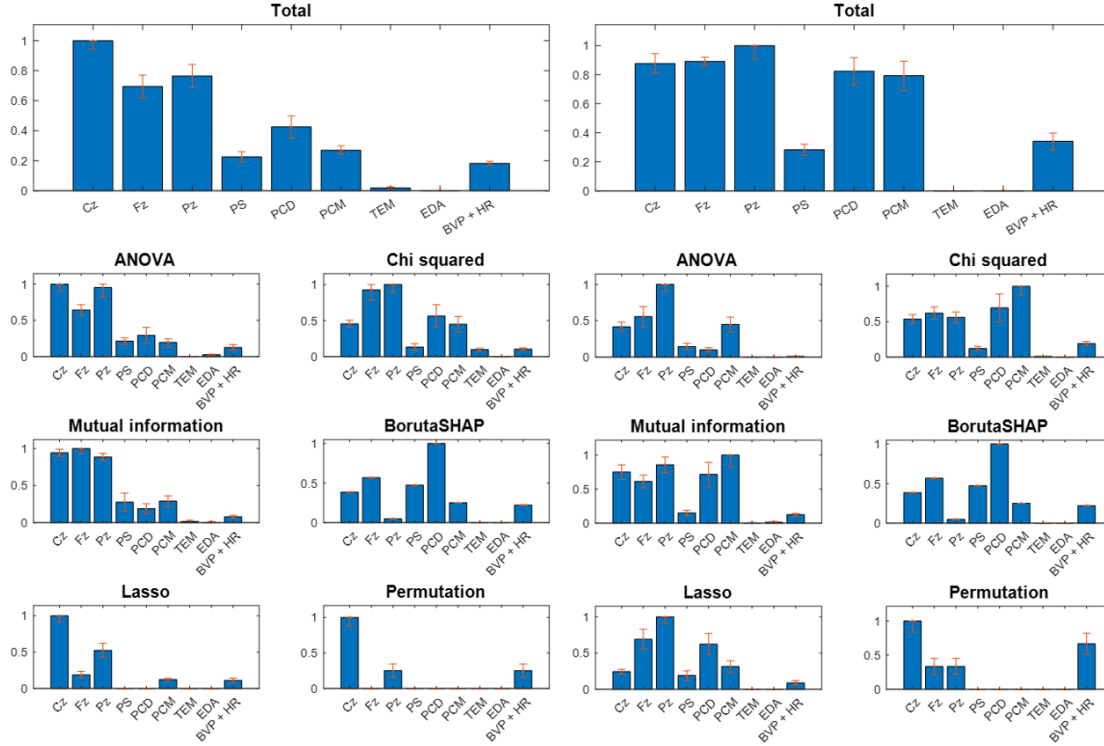


- (Left: Alerting vs Collision) Alerting: Alerting potential risks of collision with vehicle(s) in front / Collision: Collision with the vehicle(s)
- (Right: Alerting vs No Collision) Alerting: same as above / No Collision: No collision with vehicle(s) and autonomous driving is fully stopped before vehicle(s)

Figure 37 illustrates the importance of distinguishing between no collision risk (No Collision) and collision risks (Collision) with vehicles in Traffic Environment 1. In the ‘No Collision vs. Collision’ condition, the total feature importance reveals that all EEG features clearly exhibit higher importance than other features, indicating that the activities in these three brain regions are vital for distinguishing collision risks with vehicles (between collision and no collision conditions). Additionally, pupil shift features (PCM and PCD) also demonstrate notable importance, although not as significant as the EEG features. In the ‘No Collision vs. Shoulder’ condition, the feature importance results for distinguishing between no collision risk (No Collision) and potential risks in the adjacent lane due to vehicles on the shoulder (Shoulder) are presented. Compared to the results in the ‘No Collision vs. Collision’ condition, all features except for Cz show higher feature importance, and there is less discrepancy in feature importance between EEG and pupil-related features. It indicates that pupil features and EEG features are relatively well-distributed. The heightened importance of pupil features can be explained by the position of vehicles on the shoulder requiring participants to shift their gaze to verify targets between the driving lane and the shoulder. It points out that

pupil dynamics can be significant depending on where potential risks are located as users require shifting their gazes to assess these risks.

Figure 37. Feature analysis of three factors in Traffic environment 1: No collision, Collision and Shoulder

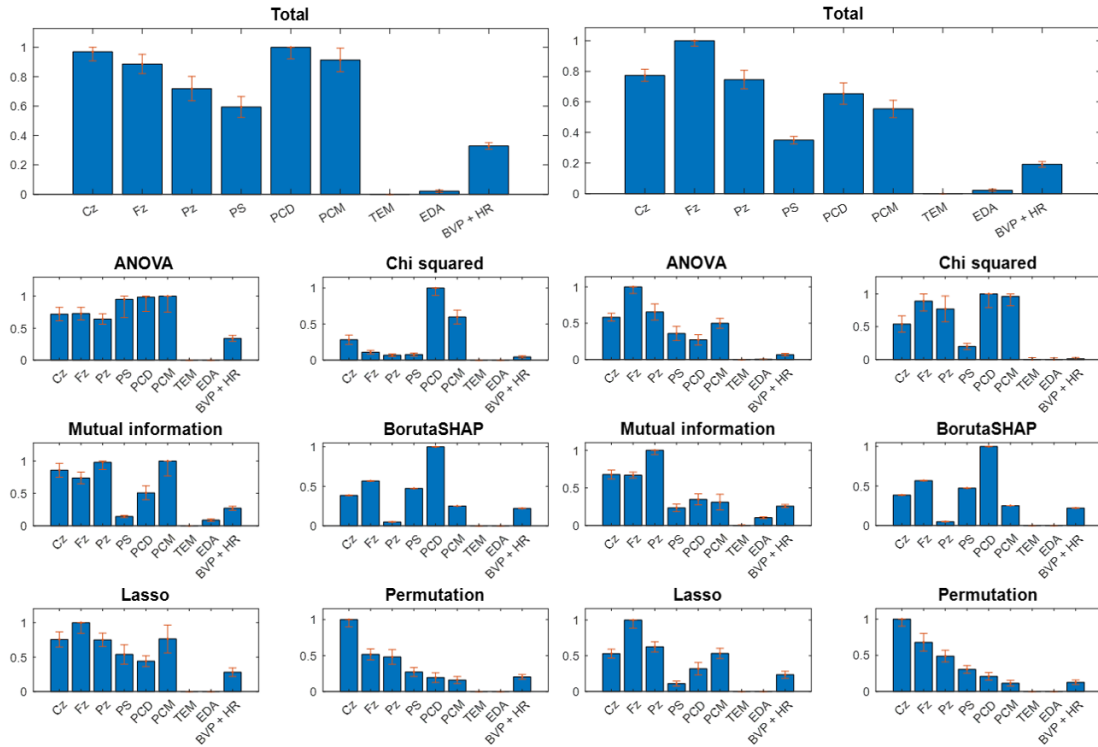


- (Left: No Collision vs Collision) No Collision: No collision with vehicle(s) in front and autonomous driving is fully stopped before the vehicle(s) / Collision: Collision with the vehicle(s).
- (Right: No Collision vs Shoulder) No Collision: same as above / Shoulder: Autonomous driving on the first lane has potential risks with stopped vehicles on shoulder.

Figure 38 illustrates the feature importance of distinguishing between alerting participants before potential collision risks (Alerting) and autonomous vehicle driving scenarios where the vehicle is aware of potential collision risks and changes lanes to avoid collisions (AV). In the ‘Alerting vs AV’ condition, the total feature importance reveals that while the average of EEG features is similar to that of pupil features, PCD (Pupil Center Degree) shows the highest importance. This heightened importance of the pupil feature, PCD, can be explained by the dynamic changes in the visual scene as the driving vehicle changes lanes to avoid collisions. In the ‘Alerting vs Avoidance by a lane change’ in self-driving scenarios, the higher feature importance observed in pupil dynamics can be interpreted as a result of participants recognizing changes in vehicle orientation due to lane changes. In contrast, in the alerting vs. collision condition, where lane changes do not occur, participants primarily perceive deceleration. The pupil responses elicited during deceleration may be less distinct than those triggered by orientation changes during lane changes, resulting

in less differentiation in alerting cases. Consequently, the importance of pupil features appears to be relatively lower in deceleration vs. alerting scenarios. Frontal EEG channels, such as Fz, are generally more sensitive to cognitive and attentional processes, including reactions to speed changes like deceleration. Studies have shown that frontal midline theta activity increases in response to collisions during motor tasks, reflecting heightened cognitive processing in these regions [227]. However, central (Cz) and parietal (Pz) channels also play significant roles in processing motor-related information. Research suggests that the frontal and parietal cortices are involved in cognitive processing during reaching movements, with theta dynamics observed in both regions [227].

Figure 38. Feature analysis of three factors in Traffic environment 2: Alerting, AV and Collision

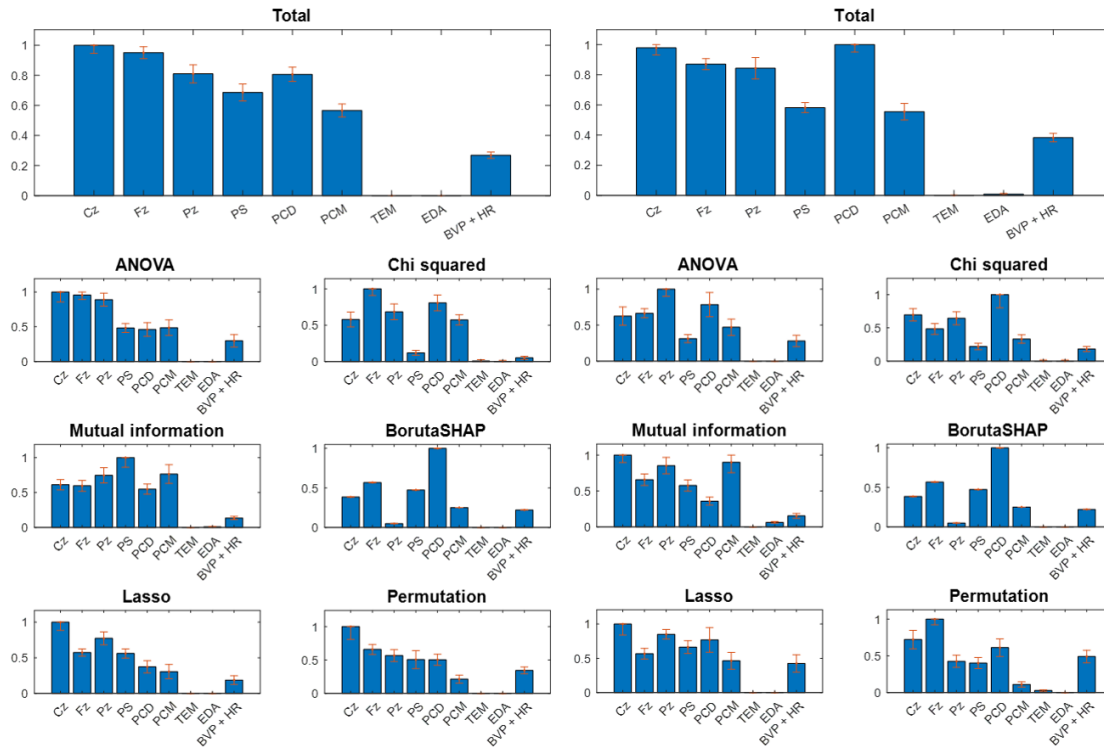


- (Left: Alerting vs AV) Alerting: Alerting potential risks of collision with vehicle(s) / AV: AV driving is fully aware of potential collision risks with vehicle(s) and changes lanes to avoid the collision.
- (Right: Alerting vs Collision) Alerting: Same as above / Collision: Collision with vehicles(s)

Figure 39 illustrates the importance of distinguishing between AV driving conditions where the vehicle is aware of potential collision risks and changes lanes to avoid collisions (AV) and the conditions where there is a collision with vehicles (Collision) in Traffic Environment 2. In the ‘AV vs Collision’ condition, the total feature importance reveals that all EEG features clearly show higher importance than other features, indicating that the activities in these three lobes are vital for distinguishing between collision and AV cases. Furthermore, the average of EEG features is higher than pupil features in both ‘AV vs Collision’ and ‘AV

vs SV2' conditions. In the 'AV vs SV2' scenario, the feature importance results for distinguishing between AV driving case and collision with vehicles at high speed (SV2) are shown. In both 'AV vs Collision' and 'AV vs SV2' conditions, feature importance rankings are the same except for the PCD feature. The only difference is shown in the PCD feature, which is of the highest importance in the 'AV vs SV2' condition. The heightened importance of the PCD feature may be explained by the dynamic changes in the visual scene at high speeds where participants need to process and react to the driving behaviour of collision avoidance. In the condition of collision without avoidance versus deceleration, the feature of change in pupil center degree (PCD) emerged as particularly important. However, when constructing a model that includes two deceleration conditions, the feature importance becomes more evenly distributed among PS, PCD, and PCM. This indicates that while PCD shows the highest importance when considering only high-speed conditions, a broader range of pupil features is taken into account when all collision scenarios are considered.

Figure 39. Feature analysis of three factors in Traffic environment 2: AV, Collision and SV2

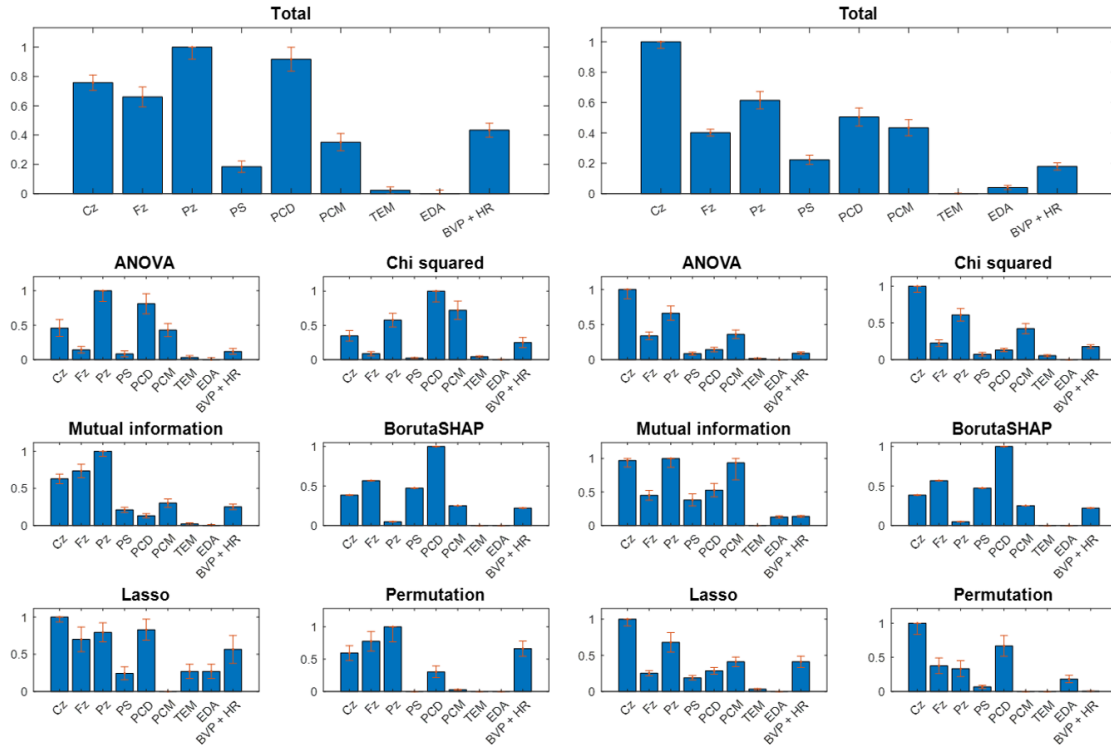


- (Left: AV vs Collision) AV: AV driving is fully aware of potential collision risks with vehicle(s) and changes lanes to avoid the collision / Collision: Collision with vehicle(s)
- (Right: AV vs SV2) AV: Same as above / SV2: Collision with vehicles(s) at speed around 85 km/h that is same as 'AV' condition when AV driving changes lanes to avoid a collision at the same speed, 85 km/h.

Figure 40 presents the importance of distinguishing between scenarios with No Collision (no collision risk) vs Alerting or AV conditions in Traffic Environment 2. In the 'No Collision vs. Alerting' condition, the

importance of the total feature highlights that EEG features are highly significant, but the PCD feature is comparable to the Pz feature, which is prioritized first. The Pz and PCD features especially exhibit the highest importance, suggesting that activity in the parietal lobe and changes in gaze direction are vital for differentiating these conditions. Furthermore, the importance of these features is comparable to the results in the same condition, ‘No Collision vs Alerting,’ with pedestrian collision risks. In this condition, the feature importance of BVP + HR exceeds that of PCD and PS features. In the ‘No Collision vs. AV’ condition, the feature importance of Cz is overwhelmingly high, which indicates that the condition can highly rely on the CZ feature to distinguish cases between no risks and AV. However, the feature importance sum of other EEG and pupil features is still higher than the Cz feature.

Figure 40. Feature analysis of three factors in Traffic environment 2: No Collision, Alerting, AV

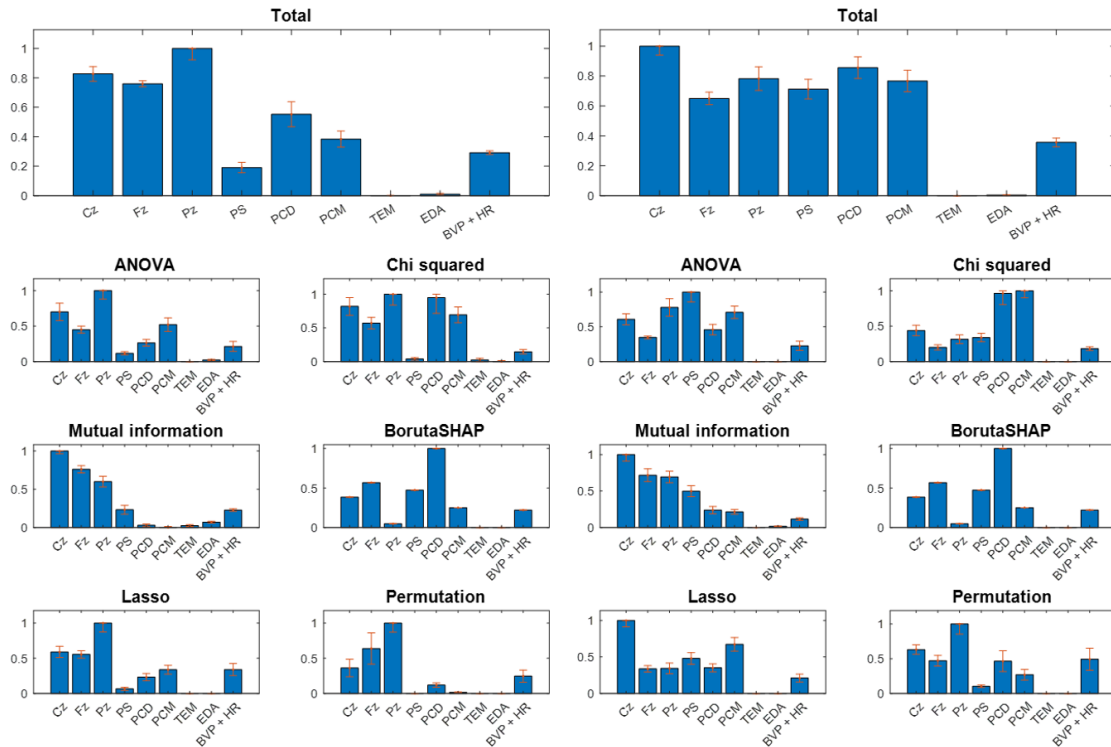


- (Left: No Collision vs Alerting) No Collision: No collision with vehicle(s) / Alerting: Alerting potential risks of collision with vehicle(s)
- (Right: No Collision vs AV) No Collision: Same as above / AV: AV driving is fully aware of potential collision risks with vehicle(s) and changes lanes to avoid the collision.

In the ‘No Collision vs. Collision’ scenario, the total feature importance reveals that EEG features show higher importance compared to pupil-related features, indicating that the brain signals are vital for distinguishing between no collision and collision conditions as shown in Figure 41. Additionally, PCD is of the highest importance among pupil features but is insignificant and less than any of the EEG features.

Compared to the same condition in collision risks with pedestrians, Cz and PCM features show less importance in collision risks with vehicles. In the ‘SV1 vs. SV2’ condition, the feature importance results for distinguishing between collisions at varying speeds are shown. Overall, EEG and pupil-related features are significant, and the differences among these features are minimal, indicating that the importance values are well distributed. These results are similar to those observed in conditions involving collisions with pedestrians, except for the importance of the PS feature. The PS feature shows heightened importance in conditions involving collisions with vehicles, which may be attributed to either the difference in the size of collision objects, the differences in the collision object type between vehicles and pedestrians, or even both.

Figure 41. Feature analysis of three factors in Traffic environment 2: No Collision, Collision and Collision severity



- (Left) No Collision: No collision with vehicle(s) and autonomous driving is fully stopped before with vehicle (s) / Collision: Collision with vehicle(s).
- (Right) Severity level 1~ SV1: Collision with vehicle(s) at vehicle speed, 50~60 km/h / Severity level 2~ SV2: Collision with vehicle(s) at a higher vehicle speed around 85 km/h than SV1.

In Traffic Environment 2, the analysis of feature importance reveals significant differences in pupil and EEG features. In the ‘No Collision vs Collision’ condition, Comparing no collision to collision conditions, the total feature importance graph shows that no collision conditions exhibit lower importance for PS, PCD, and PCM. This indicates reduced cognitive load, stress, and physiological arousal in no collision conditions,

while potential collisions reflect increased cognitive and emotional stress due to collisions. Comparing different collision severities based on speed in Traffic Environment 2, higher speeds (85 km/h) show greater feature importance across all features. This reflects increased cognitive demands, stress, and physiological arousal associated with high-speed collisions. In contrast, lower speeds (30 km/h) show reduced feature importance, indicating lower cognitive and emotional stress. This pattern highlights the heightened cognitive and physiological challenges posed by high-speed collisions, emphasizing the importance of managing speed to reduce stress and enhance safety in autonomous driving scenarios. Overall, in vehicle collision conditions, the importance of pupil features increased as the differences in vehicle speed became more pronounced when comparing various conditions, such as deceleration changes, avoidance, and no-collision scenarios. This trend aligns with findings from other studies [228, 229]. However, detailed and universally applicable patterns regarding these features were not clearly observed across the different conditions.

5.5 Discussion

The classification performance results demonstrate robust classification capabilities across various class types and granularity levels in three traffic environments. These classification conditions represent diverse factors such as traffic speeds, collision speed, types of collision objects (pedestrians and vehicles), and the number of pedestrians. Overall, we achieved superior performances for most classification conditions in our study. Binary classifications, e.g., distinguishing between collision and no collision scenarios, achieved great performance metrics consistently exceeding 0.9. Even as the number of classes increased to up to 10, there was no significant decline in performance. By Comparing the model performance with fewer classes, the results remained similarly high, consistently exceeding 0.9. However, the classification model showed more stable and higher performance in scenarios involving collision risks with pedestrians compared to those involving vehicles. The most simplified classification model that included both collision risks with pedestrians and vehicles exhibited the lowest and most unstable performance, around 0.78 in all metrics except for the ROC value. For ROC performance, most classification models showed values over 0.95, excluding the model for classes involving collisions with both vehicles and pedestrians. However, the highest granularity level contains a total of 10 classes, which may lead to overfitting issues in the user perception model when detecting non-risk and risk cases. In practical applications, such a high level of granularity may not be necessary. Therefore, by reducing the granularity of the model, for example, by regrouping the classes into fewer categories, the risk of overfitting can be mitigated. In the results of feature analysis, most factors showed high feature importance for EEG and pupil features. However, depending on the factors, the importance of specific EEG features varied. For example, Pz was the most important feature in the ‘Adult vs. Kid’ condition while Fz was the most important in the ‘Alerting vs. Collision/No Collision’

condition. These variations indicate that different factors influence which features are more important for distinguishing between classes. The influence was even more pronounced for pupil features, where the importance of PS, PCD, and PCM varied significantly depending on traffic conditions. For instance, the differences among the three pupil features were substantial in the ‘Collision vs. No Collision’ condition on a highway. In contrast, these differences were minimal in a small-town scenario, resulting in a much higher sum value of pupil feature importance. This overall increase in pupil feature importance highlights their enhanced role in these conditions. Furthermore, not all physiological features contributed significantly to the classification models across all conditions. For instance, overall, EDA and skin temperature showed minimal feature importance across most conditions. Similarly, while heart activity-related features displayed some relevance, their importance was generally not as pronounced and varied depending on the conditions. These results indicate that the most relevant features for distinguishing between classes can vary depending on the factors involved. Specifically, while certain physiological features, such as EEG and pupil-related metrics, play a critical role, others may exhibit less consistent contributions depending on different granularities. This implies that not all recording devices may be necessary based on the result of feature importance analysis, e.g., Empatica E4 (the wrist wearable type sensor containing PPG, EDA skin temperature, etc). These aspects should be carefully assessed in future studies.

5.5.1 Potential Applications of Subjectively Perceived Risks by Physiological Sensing Methods

Various objective measures rely on advanced sensing technologies, such as computer vision, LiDAR, radar, and ultrasound, in automated driving. These technologies excel in detecting objects around autonomous vehicles and aiding in the perception of various environmental factors. However, despite the advancements in sensing technologies, accident cases involving AUTOPILOT systems have been reported [230]. Specifically, even computer vision-based self-driving systems may fail to identify objects or predict their trajectories [231]. In the case of LiDAR, its performance can be adversely affected by weather conditions such as snow, rain, or fog, which impacts its ability to identify objects and avoid collisions [231]. Additionally, current standards for automated driving, such as ISO 26262 and ISO 21448, are criticized for being insufficient to fully address the safety-critical requirements of AI-based systems [232]. The newly introduced ISO PAS 8800: Road Vehicles – Safety and Artificial Intelligence, first published as a preliminary edition in December 2024, offers compliance guidance for improving the reliability and safety of autonomous driving systems using AI. However, it remains a provisional technical document and has not yet been promoted to a formal international standard (ISO Standard). These limitations highlight the need for further testing of automated driving systems under diverse environmental conditions.

Conventionally, drivers did not rely on physical sensors and instead evaluated their surroundings, hazards, and risks independently. However, past studies have shown that driver-controlled vehicle operations are prone to significant uncertainties. For example, various types of accidents caused by drivers and the psychological, physiological, and emotional aspects affecting driving performance indicate that vehicle control relying on human drivers is subject to numerous influencing factors [233, 234, 235]. Compared to automated driving systems, it is currently difficult to directly compare the number of accidents caused by human drivers versus those caused by autonomous vehicles due to the limited number of automated vehicles and their restricted operational regions. However, physical sensors demonstrate superior performance in detecting environmental information, even in scenarios where blind spots and limited driver sight are challenges and other scenarios [236, 237]. This suggests that physical sensors are safer and more reliable than human drivers in many situations. If some environmental challenges, such as fog, snow, and rain, can be overcome, the errors associated with physical sensors could be further reduced. However, there is a new challenge in automated driving systems, e.g., external risks such as cybersecurity threats (e.g., hacking) [238]. Despite these advancements, full automation remains limited due to the lack of necessary infrastructure, testing restrictions, ambiguities surrounding liability, and government regulations [239, 240]. Furthermore, whether drivers fully trust self-driving systems continues to vary significantly among individuals [241, 242, 243, 244]. This indicates that users need the ability to intervene directly when using self-driving features. Notably, under current autonomous driving standards, such as Level 3, drivers are required to intervene only when the system requests drivers' takeover. However, there are moments when drivers may choose to intervene, e.g., take over the control. These decisions often rely on subjectively perceived risk, trust or subjectively evaluated driving environments [236, 237, 245]. Such subjective judgments may differ from the risk perceived by physical sensors and automated driving systems. These differences can lead to situations where drivers attempt to take over control at times when the system does not request the takeover. Conversely, there may be instances where the system requests a takeover, but the driver determines that manual control is unnecessary. Although in Level 3 systems, drivers are obligated to comply with takeover requests from the automated system, these discrepancies highlight potential mismatches between human and system perceptions. If physiological signals can be utilized to generate and associate features related to subjective risk perception, it would enable a deeper analysis of the scenarios described above. For instance, in cases where the automated driving system does not request a takeover, but drivers initiate one, physiological data could help determine whether the decision was driven by perceived risks or hazards. It could also help identify whether the issue stemmed from sensor failures, a lack of trust or interaction between the user and the system, or differences in perception between the two. Addressing these challenges through research could reduce unnecessary takeovers by drivers and significantly improve trust between automated systems and users. This, in turn, could enhance the efficiency

and reliability of autonomous driving systems. However, achieving these goals will require extensive research into various aspects of driver-system interaction, subjective risk perception, and trust-building mechanisms. Below is a list of potential research topics to achieve these objectives.

- 1) **Understanding Subjective Risk Perception in Automated Driving:** Investigate how drivers perceive risk in different driving scenarios, such as high-speed conditions, intersections, or adverse weather and develop methods for measuring subjective risk perception, e.g., using self-reports, physiological signals (e.g., EEG, heart rate variability, eye tracking), and behavioural observations.
- 2) **Mitigating Perception Differences Between Drivers and Systems:** Identify scenarios where drivers' subjective risk perception significantly differs from the risk assessment of automated systems. Additionally, investigate cases where drivers attempt to take over control while the automated driving system continues self-driving, as well as instances where the system requests a takeover, but the driver prefers to let the system maintain control.
- 3) **Reducing Unnecessary Takeovers:** Create predictive models that identify unnecessary takeovers based on driver behaviours and physiological signals. These models can enable systems to provide information about route planning, perceived objects, hazards, and risk factors. Such system-provided information can help drivers reconsider and potentially mitigate their intentions for unnecessary takeovers.
- 4) **Longitudinal Effect and Trust-Building Mechanisms:** Investigate the longitudinal impact of the advanced automated driving system (described above) that provides additional information when drivers attempt to take over vehicle control. In addition, understanding how this system influences changes in drivers' trust in automated driving systems over time can offer valuable insights into the trust-building mechanisms between drivers and automated driving systems.

To pursue these research topics, specific areas also need to be explored in detail. In particular, studies on driver-system interaction are crucial, focusing on how the interaction will take place and what methods will be used to provide information (e.g., visual, auditory, etc.). Additionally, research on the preprocessing of physiological signals is necessary to minimize artifacts while generating features that maximize the inclusion of salient information. In addition to the aspects discussed earlier, there are other potential applications utilizing various physiological responses. These possibilities are outlined in detail below.

- 5) Recognizing a Driver's Ability to Intervene: Features from physiological signals can improve the transfer of vehicle control from the automated system to the driver.

For example, in low-risk scenarios, such as a collision risk at low speeds with road infrastructure (e.g., poles or barriers) that does not involve pedestrians or other vehicles and where no secondary risk is anticipated, or situations like navigating through a non-busy parking lot at low speeds (5–15 km/h), physiological signal features can be utilized to determine perceived risk. In such cases, the automated driving system could prompt the driver or passenger for consent to maintain control. For high-speed driving, physiological signals' potential risk perception can be expanded beyond specific moments to account for broader traffic situations. If a driver's perception of a potential risk level can be identified through physiological signals, this information can be used to assess the user's state. Similar to the low-risk scenarios, the automated driving system could provide options for the user to consent to the continued use of self-driving features.

However, the implementation of such features requires additional explorations in the following areas:

- The association between psychological states or perceived risk levels and features extracted from physiological signals.
- Studies involving diverse risk scenarios and reliable methods for measuring these risks.
- Physiological signals' susceptibility to various sensory stimuli or environmental changes (e.g., auditory distractions). The feature extraction method based on fixation epochs proposed in this study is one potential strategy. However, to utilize various signal modalities effectively, preprocessing strategies tailored to each modality must be developed.

For situations where drivers perceive relatively higher risks, two distinct cases can be identified:

- Time-sensitive control decisions: For instance, rapidly approaching collision risks where quick decisions and takeover processes are critical.
- Non-time-sensitive control decisions: Such as scenarios involving multiple students walking in school zones, where decisions are less urgent.

In the latter case, a user-consent-based approach, similar to low-risk scenarios, can be applied. However, in time-sensitive scenarios, quick decisions and immediate takeover processes are necessary. For example, in approaching collision situations, braking could be initiated before the driver takes over control, enabling simultaneous steering adjustments. Information for collision avoidance, derived from physical sensors,

could also be displayed through tools such as a head-up display, empowering the driver to respond directly to the risk. These features can be activated when risk perception is detected through physiological signals. To reduce false alarms in such systems, adding modalities such as EMG signals to capture the driver's intentions could enhance the safety and effectiveness of the takeover process. Integrating features from multiple modalities could help ensure a more reliable and secure transition of control.

- 6) **Detecting Potential Sensing Failures:** Features extracted from physiological signals may provide insights into situations where sensing technologies fail. However, there are inherent limitations to the use of physiological signals. For the applications proposed, additional research, as outlined in the previously listed areas, should be explored as part of future studies.

To elaborate further, while recording physiological signals during physical sensor malfunctions may not pose significant challenges, similar features may be generated under unrelated circumstances or stimuli. For instance, physiological signals acquired in response to stimuli like sharp or loud noises (e.g., modified vehicles or motorcycles with loud engines) or even while listening to scary stories on the radio may produce similar features. More research is necessary to determine whether these scenarios present genuine issues or are manageable within the application. It is also critical to consider that risk perception may occur due to external distractions or unrelated situations, highlighting the need to study how various sources of stimuli influence physiological signals and the responses they elicit. Additionally, individual perceived risks can vary significantly, as personal thresholds and interpretations of risk differ. A potential solution is the development of personalized models that account for these individual differences. In particular, for brain-related analysis, it is important to focus on regions of the cerebral cortex associated with specific sensory processing. A detailed examination of multiple EEG channels is needed to identify correlations between these regions and sensory stimuli. For instance, visual and auditory stimuli may elicit responses from distinct EEG channels, showing differences in specific frequencies and amplitude during various epochs [246, 247, 248]. This represents an area of untapped potential for further research into utilizing EEG signals. Moreover, when multiple stimuli are presented simultaneously, it is necessary to investigate whether preprocessing techniques can extract desired features and distinguish them from those associated with risk perception. The differences between risk-related features and other unrelated features should also be studied [249]. Using physiological signals for collision risk detection and applying them to braking systems presents challenges, particularly due to instances where physiological signals may resemble those generated in non-collision-related situations. Furthermore, physiological signals are vulnerable to artifacts, which limits the accuracy and reliability of models based solely on these signals for precise braking scenarios. Both applications discussed earlier also require extensive additional research to address these limitations.

Rather than directly applying physiological signals to some vehicle control in critical situations, it would be more practical to develop applications that utilize them for enhancing driver-system interaction, e.g., the process of takeover decisions, avoiding unnecessary takeover and safety-assured environments. Focusing on such applications, where physiological signals complement existing safety measures, would likely yield more reliable and effective results.

5.5.2 Optimizing Temporal Resolution in Critical Situation Detection: EEG, Eye-tracking, and Other Physiological Signals

Risk detection tasks in cognitive computing can be enhanced by incorporating physiological signals, thereby improving system performance and human-machine interactions. However, studies focusing on risk detection using physiological and eye-tracking signals are scarce. As discussed in Chapter 2, perceiving errors and risks in systems can be interconnected psychological processes that influence how individuals interpret and respond to their environment. Both involve cognitive evaluations: risk perception pertains to assessing potential hazards, while error perception involves recognizing mistakes or inaccuracies. These perceptions are subjective, shaped by personal experiences, knowledge, and beliefs, and significantly impact decision-making and behaviour. Additionally, temporal factors are crucial in detecting errors and risks, including processing time, the computational complexity of algorithms, and the epoch segmentation of physiological signals. In Chapter 3, we analyzed the computational complexity of our proposed algorithm using Big-O notation, demonstrating its compatibility with other EEG preprocessing algorithms employed in real-time applications. Therefore, it is essential to consider the epoch segmentation of human responses, such as EEG, eye-tracking, and other physiological signals. Most research primarily focuses on modalities like EEG and eye-tracking due to their effectiveness in detecting cognitive states and attentional shifts [7, 18, 250, 251, 252, 253]. Eye-tracking systems, in particular, are extensively used to monitor pupil changes and gaze patterns, providing valuable insights into attentional and perceptual processes during error-prone tasks [250, 254]. While multimodal approaches could offer significant advantages, studies explicitly integrating other physiological signals, e.g., ECG, PPG, GSR, and skin temperature, alongside EEG and eye-tracking for error and risk perceptions remain limited. Instead, most existing research explores these multimodal systems in broader contexts, such as emotion recognition, stress analysis, and cognitive workload assessment [255, 256, 257]. These broader contexts often lack the time-sensitive focus required for error and risk detection tasks. For example, a study utilizing PPG and GSR signals for emotion recognition identified a 3-second window as the optimal duration for signal processing [258]. Similarly, Ayata et al. [259] and Zhang et al. [260] demonstrated that ECG, PPG, and GSR achieved peak performance within a 2- to 4-second epoch window. In the case of skin conductance, studies reveal that event-related

responses typically take 2 to 6 seconds to reach their peak following stimulus onset [261], and full recovery can require 6 to 20 seconds [128]. This suggests that GSR features extracted from epochs shorter than 2 seconds or before full recovery (6–20 seconds) may not be adequate for detecting either errors or risks in visual stimuli. These findings contrast with the response times observed in EEG and pupil diameter data. For instance, Wessel et al. reported significant changes in pupil diameter and salient brain responses within 200–300 milliseconds after stimulus onset [18]. Given the time-critical nature of some detection tasks, especially those related to safety, rapid processing, from epoch extraction to feature computation, is required, often within a 1-second window. Physiological signals other than EEG and eye-tracking, such as GSR, PPG, and ECG, may lack the temporal resolution needed for such tasks. Existing studies tend to emphasize longer epoch durations, such as 2 to 4 seconds, which may not align with the time constraints of error detection in certain conditions. Instead, these physiological signals could be more suitable for applications where immediate detection is not required, or where longer epoch durations and processing delays are acceptable. Such tasks include quality control in manufacturing, forensic error detection, and data outlier detection, where delays do not lead to harm or operational failure. In these scenarios, the focus shifts toward accuracy and thorough analysis, making longer processing windows more practical and beneficial. As demonstrated in the results, features related to EEG and eye-tracking exhibited the highest importance, followed by heart rate-related features. In contrast, features derived from GSR and temperature showed lower importance. In conclusion, my work contributes to this emerging field by focusing on EEG and eye-tracking as core modalities. These modalities are well-established for their sensitivity to cognitive and attentional states, making them particularly suitable for time-sensitive tasks like error detection.

5.6 Limitations

This study explores risk detection using physiological and eye-tracking signals and proves that there is a potential of using human responses to develop risk detection models. However, there are several limitations need to be addressed to enhance the generalizability and applicability of these models. First, the small sample size limits the generalizability of the risk detection models. Future research should involve larger and more diverse participant groups to ensure consistent model performance across different user demographics. Second, while this study considers certain traffic situations with factors like vehicle speeds, collision subjects, and road types, the scenarios lack sufficient diversity to represent all traffic conditions. Therefore, various factors such as weather, a variety of vehicle speeds, deceleration rate, road types, etc, should be considered. Third, the study lacks real-world applicability, as no practical application tests were conducted, e.g., real road testing. Fourth, transitioning these models from exploratory to practical applications introduces scalability challenges, such as managing larger datasets, handling more complex scenarios, and extending durations for realistic use cases. Addressing these challenges may involve

exploring higher-level network models, which are well-suited for high-dimensional data and diverse datasets. Fifth, the study primarily focuses on risk detection in response to visual stimuli, leaving the performance of these models under multisensory environments unexplored. Future research should assess their applicability to scenarios where auditory or olfactory stimuli are presented alongside visual cues, ensuring effectiveness in complex, multisensory settings. Sixth, real-world environments pose additional challenges for EEG and eye-tracking systems due to head, body, and eye movements. Wearable designs for eye-tracking systems can help mitigate the effects of head and body motion. Although eye movement artifacts may occur, filtering algorithms can effectively address these. However, EEG systems are more sensitive to motion, and traditional wet-electrode setups are impractical for real-world use. As shown in Figure 42, newer dry-electrode EEG systems offer promising advancements, making them more portable and suitable for non-lab settings. While these developments are still emerging, ongoing improvements suggest a potential for real-world applications in the near future. Seventh, while many studies focus on binary classification, few address differentiating levels of severity. Although initial attempts to classify severity using EEG have been made, results remain inconclusive [262]. Future research should explore incorporation of multiple severity levels in risk detection models. Eighth, only one epoch duration condition was explored in this study. As discussed earlier, while there are some temporal limitations to physiological responses to visual stimuli, it may still be possible to investigate shorter durations using selected modalities such as EEG and eye-tracking. Shorter durations could be particularly beneficial for time-critical applications. Conversely, for non-time-critical applications, longer epochs might capture more detailed information about cognitive processing and user physiological responses to stimuli. Therefore, exploring varied epoch durations under appropriate constraints could be valuable for optimizing model performance and tailoring applications to specific use cases. This approach could also provide deeper insights into the temporal dynamics of physiological responses and improve the adaptability of classification models. Lastly, there is an issue of ground truth values for risk severity levels. Various severity levels from multiple factors such as collision subjects, the number of pedestrians, and vehicle speeds at collisions may be seen obviously depending on combinations of the factors. However, risk severity levels rated by participants using survey responses or by expert annotations could work as ground truth values for the levels. Addressing these limitations is essential for advancing the utility and robustness of risk detection models in real-world applications, ensuring their scalability, adaptability, and effectiveness.



Figure 42. EEG systems with improved convenience: MindWave from NeuroSky, BR8+ from BRI and EPOC from Emotiv (Images [263]).

5.7 Conclusion

Three research questions, 3 to 5 in Section 2.5, are fully addressed in this study. The study demonstrated strong classification performance across various traffic environments considering different factors such as traffic speeds, collision types, and the number of pedestrians involved. The binary classification tasks, e.g., distinguishing between collision and no collision conditions, achieved great performances exceeding 0.9 across all metrics. Even with an increased number of classes up to 10 (RQ4), the model consistently maintained high performance, achieving metrics of over 0.9 (RQ5). However, the model showed better stability and performance in scenarios involving pedestrian collision risks than those involving vehicles. The simplest classification model that included both pedestrian and vehicle collision risks showed the lowest and least stable performance, around 0.78 in most metrics, though ROC values for most models exceeded 0.95. The exception was the model involving both pedestrian and vehicle collisions. Our findings indicate that the most relevant features for classification can vary depending on the specific factors and conditions involved in the traffic environment.

Chapter 6. Workload Estimation Models Using Physiological and Eye-Tracking Signals

Monitoring workload is essential in fields such as workplace safety, healthcare, and human-computer interaction. Recent technological advancements have facilitated the use of non-invasive physiological measures (EEG, eye-tracking, heart rate, BVP, GSR, and body temperature) to effectively assess workload states. In studies focusing on workload prediction models that utilize these physiological signals, mental workload changes are often measured by classifying workload states based on task performance, such as in the Stroop task and n-back task. Additionally, subjective assessment tools like the NASA Task Load Index (NASA-TLX) [53] are employed alongside these tasks to provide a comprehensive evaluation of workload. In our research, we aimed to develop a workload prediction model by extracting features related to workload from the data obtained in various experiments presented in this thesis.

6.1 Experiment Design

In this chapter, the eye-tracking and physiological data utilized are derived from previous experiments, specifically those related to error detection and user preference. As the detailed experimental procedures have been previously documented, they will be omitted here. Instead, the focus will be on the survey data used to measure mental workload. For this purpose, the NASA-TLX was employed. It is a widely recognized and extensively used questionnaire for assessing various factors in workload.

6.1.1 NASA Task Load Index (NASA-TLX)

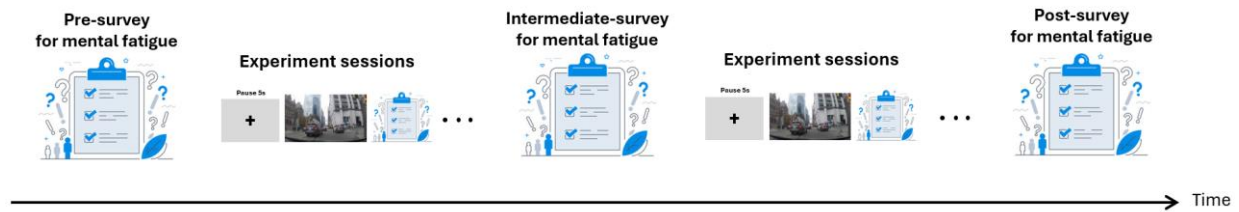
The NASA-TLX is an advanced instrument developed by the Human Performance Group at NASA Ames Research Center in the 1980s, specifically designed to assess perceived workload across a range of human-machine interaction tasks. This tool evaluates subjective workload by examining six key dimensions: mental demand, physical demand, temporal demand, performance, effort and frustration. These dimensions collectively assess the cognitive and physical requirements of the task, the time constraints, the participant's self-assessed performance, the overall effort needed, and the stress levels experienced. Participants rate each dimension on a scale ranging from 1 (Low) to 20 (High). Additionally, the importance of each dimension is determined through pairwise comparisons that lead to weighted scores that contribute to an overall workload index. For example, a participant might assign a high rating to mental demand and also give it significant weight, resulting in a correspondingly higher contribution to the overall workload score. The NASA-TLX has been widely adopted across various domains, e.g., the aviation domain, where it is used to evaluate pilot workload; healthcare for assessing stress in high-pressure environments such as

surgery; the automotive industry for analyzing driver workload and military settings to gauge the workload experienced by soldiers. Its multidimensional framework, adaptability and emphasis on subjective experience make it a critical tool for understanding and optimizing human performance, improving safety and guiding the design of user-centred systems.

6.1.2 Data Collection

In this chapter, the eye-tracking and physiological data utilized were obtained from previous experiments. However, these data alone are insufficient for establishing a baseline to assess the level of mental workload, making it impossible to develop a predictive model for mental workload. Therefore, mental workload data were collected using the NASA-TLX questionnaire as previously described. As illustrated in Figure 43, the questionnaire data were acquired at three different times for each participant: before, during and after the experiment corresponding to the pre-, intermediate- and post-survey questionnaires.

Figure 43. A schematic diagram of data acquisition for participant fatigue states.



6.1.3 Participants

In this study, we used data collected from two other studies, the CARLA experiment and the user preference studies in this thesis. Twenty-two healthy adults with normal or corrected-to-normal vision (16 men and 6 women) residing in Waterloo and Toronto (Ontario, Canada) were recruited for this experiment. The participants' ages ranged from 28 to 34 years, with an average age of 30.4 ($SD \pm 2.7$). All participants had no history of ocular or neurological diseases, disorders, or surgeries. Before participation, each participant provided informed consent either through written documentation or verbally, in accordance with the protocols approved by the Office of Research Ethics at the University of Waterloo (Applications No. 43247 and No. 43926). Existing studies have demonstrated strong and reliable responses in features derived from physiological and eye-tracking signals, effectively differentiating the intensity of cognitive demands, including task complexity and mental effort [264, 265]. For instance, pupil dilation has been shown to increase with higher cognitive load during tasks requiring significant mental effort. These findings indicate that the experimental design and physiological responses in this study are capable of producing a medium or large effect size ($d \geq 0.8$). According to power analysis, a medium effect size ($d=0.5$) requires

approximately 21 participants in a within-subject design to achieve 80% power at $\alpha = 0.05$, and a large effect size requires fewer participants than a medium effect size [266]. The number of participants required under these conditions aligns closely with the sample size recruited for this study. Additionally, existing studies on cognitive workload estimation using physiological and eye-tracking signals have recruited between 13 and 24 participants [264, 265, 267, 268]. Accordingly, the recruitment of 22 participants in this study represents an appropriate sample size for detecting statistically meaningful effects. However, to further improve the generalizability of the model, future studies should consider increasing the sample size to encompass a more diverse population. This would help account for individual variability in physiological and eye-tracking responses, ensuring that the findings can be more broadly applied across different user groups and task conditions. Expanding the participant pool would also support more robust model training and validation, which are critical for applications in complex, real-world scenarios.

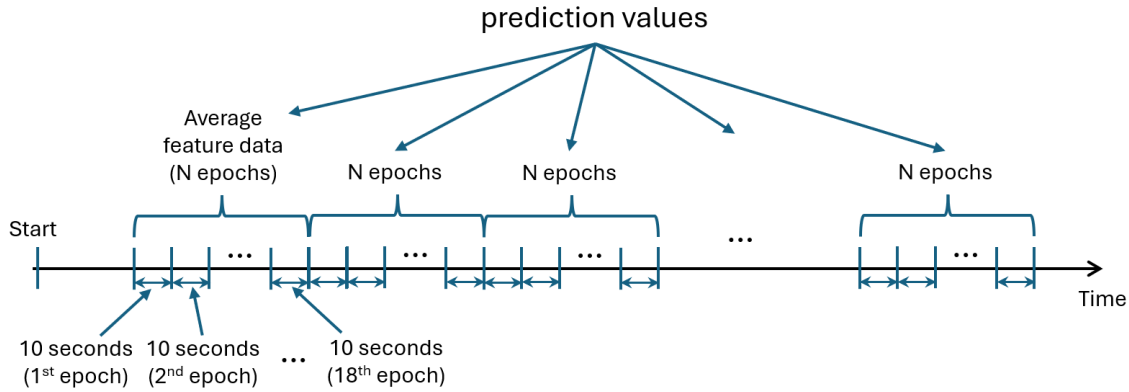
6.2 Data Analysis

As previously mentioned, the data analysis in this chapter was conducted using data obtained from prior experiments. A significant number of sessions would typically be required to build a reliable model of mental workload based on responses collected during the pre-, intermediate- and post-experiment phases. However, with only 1–3 sessions per participant, it is challenging to develop a real-time model. Increasing the frequency of surveys to compensate for this limitation could introduce errors in survey responses due to the repetitive nature of the questions over short intervals (additional references are needed to support this limitation). Measuring and quantifying the degree of fatigue, especially in states that are not highly fatigued (e.g., based on blink duration, eyelid closure, or enhanced power in certain brain frequencies), is highly challenging. It could be too late to detect such symptoms because accident risks can significantly increase in a transition period from mild to severe fatigue. As noted earlier, without sufficient NASA-TLX response data, building a mental workload prediction model using supervised learning methods is difficult. Despite these limitations, efforts were made to develop a predictive model that can account for variations over time. A variety of approaches were explored, ranging from supervised learning to semi-supervised and unsupervised learning methods. However, if we assume that mental workload levels tend to increase over time, the time variable could be transformed into class data, which allows for the construction of a regression model. Theoretically, as the total experiment duration increases over time, the mental workload level is likely to rise, given that no other activities or stimuli are provided to refresh the participants' mental workload states apart from the breaks taken before each trial begins.

6.2.1 Feature Epoch

In this study, we aim to develop a real-time workload prediction model that performs predictions on a minute-by-minute basis rather than on a second-by-second basis. Predicting at the second level may be influenced by many factors within the mental workload that may potentially affect the classification or regression data for each epoch. Therefore, we aimed to establish a robust and reliable model based on minute-level predictions. While this model does not offer real-time second-by-second predictions, it can detect workload trends and heightened workload levels within a user-defined range of 1 to 10 minutes and provide warnings accordingly. For this purpose, we collected epochs at 10-second intervals as illustrated in Figure 44 and adjusted the number of epochs according to the user-defined prediction duration to obtain a single prediction value. In this chapter, we first built predictive models based on 3 minutes.

Figure 44. Epoch extraction based on prediction duration



- N is adjustable by a user setting for a prediction duration.

6.2.2 Feature Type

In this study, the data used to build the machine learning model include questionnaire responses from the NASA-TLX as well as physiological signals. Various features were extracted from these data sources. Detailed information on these features can be found in Chapter 3.

6.3 Machine Learning Approaches to Building Workload Prediction Model

In this chapter, three different learning methods (supervised, unsupervised and semi-supervised) were employed to build the workload prediction model. For supervised learning methods, it is crucial to have sufficient class data or predictor variables to effectively develop a prediction model, whether it be for classification or regression. However, acquiring an adequate amount of class data and continuous scale data

related to workload through experiments presents significant challenges. To address this limitation, we constructed a prediction model based on the assumption that workload is likely to increase as the experiment progresses using the duration of the experiment as a predictor variable. This assumption is supported by data from the NASA-TLX. Unsupervised and semi-supervised learning methods were also employed to further refine the workload prediction model. The following sections provide detailed explanations of the ML approaches used in this process.

6.3.1 Supervised Learning

Supervised learning is a cornerstone of ML, where models are trained using labelled datasets. Each training example consists of input-output pairs with the output serving as the ‘supervision’ for the learning process. The objective is to learn a mapping from inputs to outputs that can accurately predict the output. Tasks in supervised learning include classification to predict discrete labels and regression to predict continuous values. Algorithms such as linear regression, support vector machines, decision trees and neural networks are frequently used. The strength of supervised learning lies in its ability to leverage labelled data to achieve high accuracy and reliability in predictions, making it ideal for applications where historical data is abundant and well-annotated.

6.3.1.1 Supervised Learning Algorithm: Regression

We attempt to build regression models by using the median value of the measurement time as the prediction value based on the features recorded during the experiment. The following are the eight regression models tested in this study.

Linear regression

Linear regression is a fundamental regression technique that models the relationship between a dependent variable y , and one or more independent variables X using a linear equation. The general form of the linear regression model is:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p + \epsilon$$

where y is the dependent variable and β_0 is the intercept. $\beta_1, \beta_2, \dots, \beta_p$ are the coefficients of the independent variables $x_1, x_2, \dots, x_p$. ϵ is the error term.

The coefficients are typically estimated using the Ordinary Least Squares (OLS) method, which minimizes the sum of the squared residuals:

$$\min_{\beta} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

where $\hat{y}_i$ is the predicted value of y_i .

Regression Tree

A regression tree is a decision tree algorithm used for predicting continuous values. It recursively partitions the data into subsets based on feature values to minimize the variance within each subset. The model can be represented as:

$$f(x) = \sum_{m=1}^M c_m I(x \in R_m)$$

where $f(x)$ is the prediction function and M is the number of terminal nodes (leaves). c_m is the predicted value for the region R_m . $I(\cdot)$ is an indicator function.

The splitting process continues until a stopping criterion is met, such as samples per leaf or a maximum tree depth.

Bagging Tree

Bagging (Bootstrap Aggregating) is an ensemble method that improves the stability and accuracy of machine learning algorithms. In regression, it involves training multiple regression trees on different subsets of the data and averaging their predictions. The formula for a Bagging tree is:

$$\hat{f}(x) = \frac{1}{B} \sum_{b=1}^B f_b(x)$$

where B is the number of bootstrap samples and $f_b(x)$ is the prediction from the b -th regression tree.

Each tree is trained on a bootstrap sample, which is a random sample replaced by the original dataset.

Boosting Tree

Boosting is an ensemble technique that combines the predictions of several weak learners to form a strong learner. For regression, one standard boosting algorithm is Gradient Boosting. The prediction function is:

$$\hat{f}(x) = \sum_{m=1}^M \eta h_m(x)$$

where M is the number of trees and η is the learning rate. $h_m(x)$ is the m -th weak learner (tree).

Each new tree is trained to correct the errors of the combined ensemble.

Support Vector Regression

Support Vector Regression (SVR) extends Support Vector Machines (SVM) to regression tasks. The objective is to find a function $f(x)$ that has at most ϵ deviation from the actual targets for all training data and is as flat as possible. The formulation is:

$$\min_{w,b,\xi,\xi^*} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^n (\xi_i + \xi_i^*)$$

subject to:

$$y_i - (w \cdot x_i + b) \leq \epsilon + \xi_i$$

$$(w \cdot x_i + b) - y_i \leq \epsilon + \xi_i^*$$

$$\xi_i, \xi_i^* \geq 0$$

where w and b are parameters of the model. ξ_i and ξ_i^* are slack variables. C is the regularization parameter.

Gaussian Process Regression (GPR)

Gaussian Process Regression (GPR) is a non-parametric, Bayesian approach to regression. It defines a prior over functions, which results in a posterior distribution over functions. The key components are as follows.

The mean function $m(x)$ and the covariance function $k(x, x')$ are:

$$m(x) = \mathbb{E}[f(x)]$$

$$k(x, x') = \mathbb{E}[(f(x) - m(x))(f(x') - m(x'))]$$

Given a Gaussian process prior, the prediction $\hat{y}$ for a new input x , is:

$$\hat{y} \mid X, y, x_* \sim \mathcal{N}(\mu_*, \sigma_*^2)$$

$$\mu_* = K(x_*, X)[K(X, X) + \sigma_n^2 I]^{-1} y$$

$$\sigma_*^2 = k(x_*, x_*) - K(x_*, X)[K(X, X) + \sigma_n^2 I]^{-1} K(X, x_*)$$

Neural Network Regression

Neural networks can be used for regression by configuring the output layer to have a linear activation function suitable for continuous output. A neural network consists of layers of interconnected nodes (neurons) where each connection has a weight. The general form of a neural network is:

$$\hat{y} = f(W_n \cdot f(W_{n-1} \cdot \dots \cdot f(W_1 \cdot x + b_1) + b_{n-1}) + b_n)$$

where x is the input vector. W_i and b_i are weights and biases for the i -th layer. f is the activation function (e.g., ReLU, sigmoid), and $\hat{y}$ is the predicted output.

XGBoost Regression

XGBoost (eXtreme Gradient Boosting) is an efficient and scalable implementation of gradient boosting. It builds an ensemble of decision trees in a sequential manner where each tree attempts to correct the errors of its predecessors. The objective function combines a loss function and a regularization term:

$$\mathcal{L}(\theta) = \sum_{i=1}^n l(y_i, \hat{y}_i^{(t)}) + \sum_{k=1}^t \Omega(f_k)$$

where l is the loss function (e.g., squared error) and $\hat{y}_i^{(t)}$ is the prediction at iteration t . Ω is the regularization term for the complexity of the model.

6.3.1.2 Performance Evaluation of Regression Models

Evaluating the performance of regression models in supervised learning is crucial to ensure their accuracy and reliability in predicting continuous outcomes. Three primary metrics commonly used for this purpose are Mean Squared Error (MSE), Mean Absolute Error (MAE), and the Coefficient of Determination (R^2). These metrics provide a comprehensive evaluation of regression model performance balancing the precision of prediction, MSE and MAE, and the explanatory power, R^2 .

Mean Squared Error (MSE)

MSE measures the average of the squares of the errors, where error is the difference between the actual and predicted values. It is calculated as

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

They provided a sense of the model's prediction accuracy, with larger errors being penalized more heavily.

Mean Absolute Error (MAE)

MAE, on the other hand, calculates the average absolute differences between predicted and actual values, expressed as

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$$

This metric gives a more straightforward interpretation of prediction error by considering the magnitude of the errors without squaring them, making it less sensitive to outliers compared to MSE.

Coefficient of Determination (R^2)

R^2 , or the Coefficient of Determination, assesses the proportion of the variance in the dependent variable that is predictable from the independent variables. It is calculated as

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

where $\bar{y}$ is the mean of the actual values. An R^2 value close to 1 indicates that a large proportion of the variance in the dependent variable is accounted for by the model, reflecting a good fit.

6.3.2 Unsupervised Learning

Unsupervised learning differs fundamentally from supervised learning as it deals with unlabeled data. The primary goal is to uncover hidden patterns or intrinsic structures within the data without predefined labels. Common unsupervised learning tasks include clustering where data points are grouped based on similarity and dimensionality reduction which simplifies data by reducing the number of features while retaining essential information. Algorithms such as k-means clustering, hierarchical clustering, and principal component analysis (PCA) are widely used. Unsupervised learning is beneficial in exploratory data analysis and situations where labelling data is impractical or impossible. It enables the discovery of natural groupings and associations within data, providing valuable insights and aiding in data-driven decision-making.

6.3.2.1 Unsupervised Learning Algorithm: Clustering

K-means clustering

K-means is a popular unsupervised learning algorithm used for clustering data into k -distinct groups based on feature similarity. The algorithm aims to partition the dataset into k clusters in which each data point belongs to the cluster with the nearest mean serving as a prototype of the cluster. The process begins by randomly initializing k centroids. Each data point is then assigned to the nearest centroid based on the Euclidean distance. The centroids are recalculated as the mean of all points in the respective clusters. This process of assignment and centroid updating is repeated iteratively until convergence, typically when the assignments no longer change or the centroids stabilize. Formally, the objective is to minimize the within-cluster sum of squares (WCSS), defined as:

$$J = \sum_{i=1}^k \sum_{x \in C_i} \|x - \mu_i\|^2$$

where $\|x - \mu_i\|^2$ is the squared Euclidean distance between a data point x cluster C_i .

Hierarchical clustering

Hierarchical clustering is another unsupervised learning method used to build a hierarchy of clusters. It can be categorized into two types: agglomerative (bottom-up) and divisive (top-down). Agglomerative clustering starts with each data point as a single cluster and repeatedly merges the closest pairs of clusters until a single cluster remains or a desired number of clusters is achieved. Divisive clustering, conversely, begins with the whole dataset in one cluster and recursively splits it into smaller clusters. The distance between clusters can be measured in several ways, such as single linkage (minimum distance), complete linkage (maximum distance) and average linkage (mean distance). The process can be visualized using a dendrogram that shows the nested grouping of clusters and their similarity levels.

Density-Based Spatial Clustering of Applications with Noise (DBSCAN)

DBSCAN is a density-based clustering algorithm that can find arbitrarily shaped clusters and identify noise or outliers in the data. Unlike k-means and hierarchical clustering, DBSCAN does not require specifying the number of clusters beforehand. It relies on two parameters: ϵ (the maximum distance between two points to be considered neighbours) and minPts (the minimum number of points required to form a dense region). The algorithm starts by selecting an arbitrary point and retrieving its ϵ -neighborhood. If the neighbourhood contains at least minPts points, a cluster is started. This cluster is expanded by recursively adding all density-reachable points. Points that do not belong to any cluster are considered noise. Formally,

a point p is a core point if at least minPts points are within distance ϵ of p . A point q is directly density-reachable from p if q is within ϵ of p . The process continues until all points are visited.

Gaussian Mixture Model (GMM)

The Gaussian Mixture Model (GMM) is a probabilistic model that assumes the data is generated from a mixture of several Gaussian distributions with unknown parameters. GMM is particularly useful for clustering when the clusters are not well separated and have different shapes. Each component in the mixture is a Gaussian distribution defined by its mean and covariance matrix.

The probability density function of a GMM is given by:

$$p(x) = \sum_{i=1}^k \pi_i \mathcal{N}(x | \mu_i, \Sigma_i)$$

where k is the number of Gaussian components, and π_i is the weight of the i -th Gaussian component. $\mathcal{N}(x | \mu_i, \Sigma_i)$ is the Gaussian distribution with mean μ_i and covariance matrix Σ_i .

The parameters of the GMM are estimated using the Expectation-Maximization (EM) algorithm, which iteratively performs two steps:

- 1) Expectation (E-step): Calculate the posterior probabilities (responsibilities) that each point belongs to each Gaussian component.
- 2) Maximization (M-step): Update the parameters (means, covariances and weights) to maximize the likelihood of the data given these responsibilities.

By modelling the data as a mixture of Gaussians, GMM can effectively handle clusters of varying shapes and densities, providing a flexible approach to clustering.

6.3.2.2 Performance Evaluation of Clustering Methods

Silhouette score

The Silhouette score is a measure of how similar an object is to its own cluster compared to other clusters. It provides an indication of the quality of the clustering by considering both cohesion and separation. For each data point i , the Silhouette score $s(i)$ is calculated using the mean intra-cluster distance $a(i)$ and the mean nearest-cluster distance $b(i)$. The formula is:

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}$$

where $a(i)$ is the average distance between i and all other points in the cluster, and $b(i)$ is the average distance between i and all points in the nearest c belong.

The Silhouette score ranges from -1 to 1:

- 1) A score close to 1 indicates that the data point is well-clustered and far from neighbouring clusters.
- 2) A score close to 0 indicates that the data point is on or very close to the decision boundary between two neighbouring clusters.
- 3) A score close to -1 indicates that the data point has been assigned to the wrong cluster.

The overall Silhouette score for a clustering solution is the average Silhouette score of all data points, providing an evaluation of the clustering quality.

Davis-Bouldin index

The Davies-Bouldin Index (DBI) is a metric for evaluating clustering algorithms focusing on both the intra-cluster similarity and inter-cluster differences. Lower values of the DBI indicate better clustering. The DBI is calculated as follows:

$$DBI = \frac{1}{k} \sum_{i=1}^k \max_{j \neq i} \left(\frac{S_i + S_j}{D_{ij}} \right)$$

where k is the number of clusters. S_i is the average distance between each point in the cluster i and the centroid measure of intra-cluster dispersion). D_{ij} is the distance between the centroids of clusters i and j .

The index measures the average similarity ratio of each cluster with its most similar cluster, balancing the trade-off between the cluster compactness (intra-cluster dispersion) and the separation (inter-cluster distance).

Calinski-Harabasz index

The Calinski-Harabasz Index, also known as the Variance Ratio Criterion, evaluates the quality of clustering by examining the ratio of the sum of between-cluster dispersion and within-cluster dispersion for

all clusters. Higher values of the Calinski-Harabasz Index indicate better-defined clusters. The index is defined as:

$$CH = \frac{\text{Tr}(B_k)/(k-1)}{\text{Tr}(W_k)/(n-k)}$$

where $\text{Tr}(B_k)$ is the trace of the between-cluster dispersion matrix. $\text{Tr}(W_k)$ is the trace of the within-cluster dispersion matrix. k is the number of clusters, and n is the total number of data points.

The between-cluster dispersion matrix B_k measures the dispersion of the cluster centroids relative to the overall data means. In contrast, the within-cluster dispersion matrix W_k measures the dispersion of data points within each cluster. The Calinski-Harabasz Index evaluates how well the clusters are separated and how compact they are, which provides a robust measure of clustering performance.

6.3.3 Semi-Supervised Learning

6.3.3.1 Semi-Supervised Learning Methods

Semi-supervised learning bridges the gap between supervised and unsupervised learning by utilizing both labelled and unlabeled data. This approach is particularly advantageous when labelled data is scarce or expensive to obtain but a large amount of unlabeled data is available. Semi-supervised learning algorithms aim to improve learning accuracy by leveraging the abundant unlabeled data to better understand the data distribution and enhance model training. Techniques such as self-training, co-training, and graph-based methods are commonly employed. The semi-supervised approach strikes a balance as it provides the benefits of supervised learning's accuracy with the flexibility of unsupervised learning, making it suitable for real-world applications where complete labelling is not feasible.

6.3.3.2 Performance Evaluation of Semi-Supervised Learning Methods

Based on semi-supervised learning methods, each feature row was assigned to a particular class of workload levels according to our class decision process. This process was informed by both unsupervised learning and the classification criteria derived from the NASA-TLX. Detailed information on this approach can be found in Section 4.5. The workload classification model was then built using supervised learning methods. The model performance was evaluated using conventional performance metrics such as precision, recall, F1-score, accuracy and ROC.

6.4 Results: NASA-TLX and Performance of Machine Learning Approaches

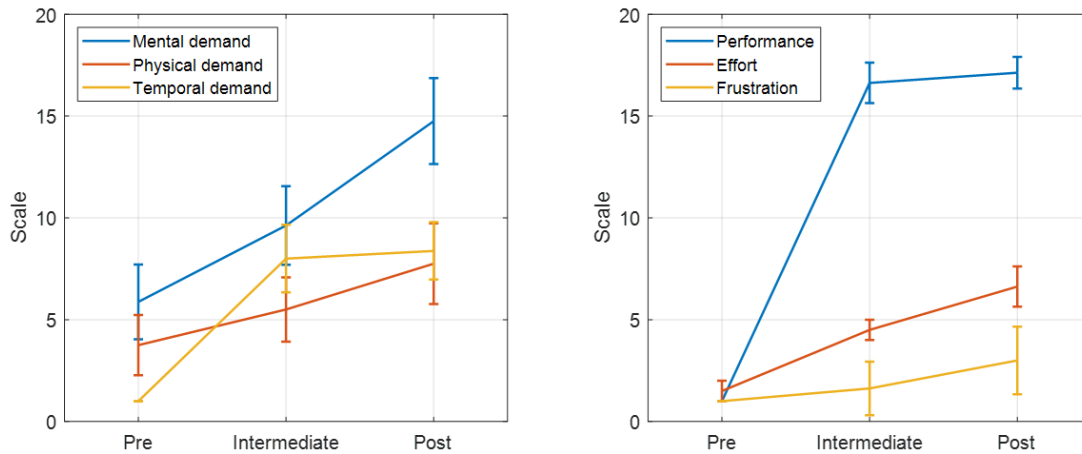
This chapter developed the workload prediction model using NASA-TLX results in conjunction with three learning methods: supervised, unsupervised (clustering) and semi-supervised (designed label decision).

6.4.1 NASA-TLX Result

We analyze changes in mental workload scores based on responses to the NASA-TLX questionnaire. These are collected during experiments, pre-, intermediate- and post-experiment. As our experiments did not require physical activities, a few dimensions in NASA-TLX, such as physical demand and frustration, may be less important than some dimensions related to tasks, e.g., mental and temporal demands. In Figure 45, the six dimensions of NASA-TLX at three stages (pre-, intermediate- and post-experiment) are shown. For mental demand, there is a steady increase from pre- to intermediate- to post-experiment, indicating that the cognitive effort required by participants progressively grew over time. This trend may appear unexpected in an experiment involving repetitive tasks, such as watching driving videos, as workload is typically assumed to remain constant or even decrease due to task familiarity. Additionally, during the pre-session phase, participants were instructed to respond to the NASA-TLX survey based on their current state, specifically reflecting their present mental fatigue or perceived demands before the session. This approach likely prompted participants to evaluate their responses with consideration of their mental fatigue levels prior to the experiment. As the experiment progressed, the intermediate-session responses appeared to have reflected the physical and mental demands experienced during the first hour of the study. These results seem to capture the overall demands of the session rather than the workload of a single task. Similarly, in the post-session phase, the NASA-TLX responses seem to represent the cumulative mental and physical demands experienced throughout the entire two-hour session. This suggests that participants interpreted the workload not as task-specific but as an evaluation of fatigue experienced during the full session. This interpretation can align with findings from previous studies, which suggest that repetitive tasks and sustained attention can result in elevated mental workload and increased mental fatigue [269, 270, 271]. However, these findings may not align with the results of other cognitive workload research conducted under conditions involving simple and repetitive tasks, where mental workload was measured using self-assessment tools. It is because mental workload can be influenced by various factors [46]. Physical demand, like mental demand, was also self-evaluated before the experiment started. The results show a slight increase at the intermediate stage and a further rise post-experiment, suggesting growing physical strain during the experiment. However, the scale magnitude for physical demand is smaller than that of mental demand, approximately half the scale of the mental demand. Temporal demand increases from the initial stage to the intermediate and then stabilizes, reflecting an initial adjustment to time pressure after the

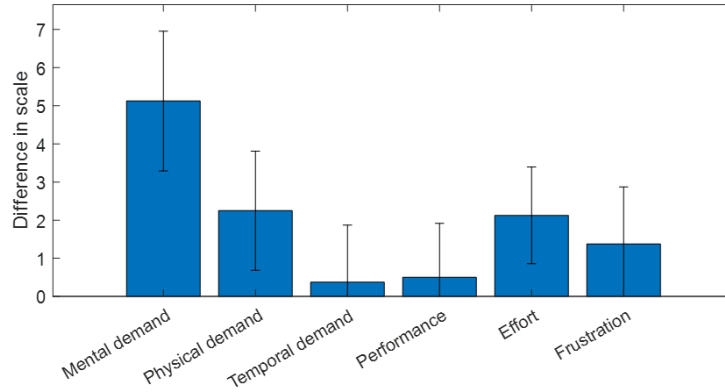
experiment starts. There are no significant changes in temporal aspects during the experiment. Additionally, the low scale values indicate that the experiments were not designed to induce time pressure. For other dimensions such as performance, effort and frustration, the initial responses are less reliable as they were recorded before the experiment began. These dimensions show slight increases from the intermediate to the post-experiment stages; however, the increases are not as significant compared to those in mental and physical demands. Overall, mental demand is highly prioritized, followed by the relative importance of physical and temporal demands and effort. As shown in Figure 46, the scale differences between the intermediate- and post-experiment NASA-TLX dimensions indicate that mental demand exhibited the most significant increase. This suggests a notable rise in mental demand as the experiment progressed. The difference in physical demand was relatively low, which aligns with the fact that the experiments were designed with minimal physical activity and demand. Furthermore, little to no difference was observed in temporal demand and performance, while differences in effort and frustration were also minimal. This suggests that the difficulty level of the tasks remained consistent throughout the experiment and that the tasks were not inherently challenging. Additionally, the lack of significant change in frustration indicates that the tasks had little to no impact on frustration levels.

Figure 45. Six dimensions of NASA-TLX in pre-, intermediate- and post-experiment



Therefore, we decided to determine the weights for each dimension to calculate the total workload. While it is possible to assign weights simply by ranking the dimensions, the average scale value for each dimension varies across different stages. Hence, we opted to determine the weights using the following equation which considers both the average and standard deviation.

Figure 46. The difference in each dimension between intermediate- and post-experiment NASA-TLX.



$$\mu_i = \frac{\sum (scale_{i,post} - scale_{i,intermediate})}{N}$$

$$\sigma_i = \sqrt{\frac{\sum ((scale_{i,post} - scale_{i,intermediate}) - \mu_i)^2}{N - 1}}$$

i = Dimension in NASA TLX, N = Number of participants

Scale of each dimension at intermediate stage (Experiment) = $scale_{i,post}$

Scale of each dimension at post stage (Experiment) = $scale_{i,post}$

First, we determined the scale differences between the intermediate- and post-experiment NASA-TLX responses for each participant to calculate the weight for each dimension. By using the equations provided above, the weights for each dimension were calculated based on the mean (μ_i) and standard deviation (σ_i) of these scale differences as shown in the equation below.

$$Weight_i = \left[\frac{\mu_i / \sigma_i}{\sum_i \mu_i / \sigma_i} \right] \times \left[\frac{\mu_i}{\sum_i \mu_i / N_{dimensions}} \right]$$

$N_{dimensions}$ = number of dimensions (6 dimensions in NASA TLX)

The equation consists of two terms. In the first term, the mean value (μ_i) of each dimension for the participants is divided by the standard deviation (σ_i) of the same dimension. The sum of these values across all six dimensions is then used to normalize the result of the first term. The second term adjusts the weights based on how each dimension's mean (μ_i) compared to the overall average mean across all dimensions. If a dimension's mean is higher than the overall average, its weight increases; conversely, if it is lower, the

weight decreases. Ultimately, the product of these two terms gives the weight value of each dimension, $Weight_i$. Using the weights calculated from this equation, we can compute the total workload for a participant as shown below. Table 15 presents the calculated weights.

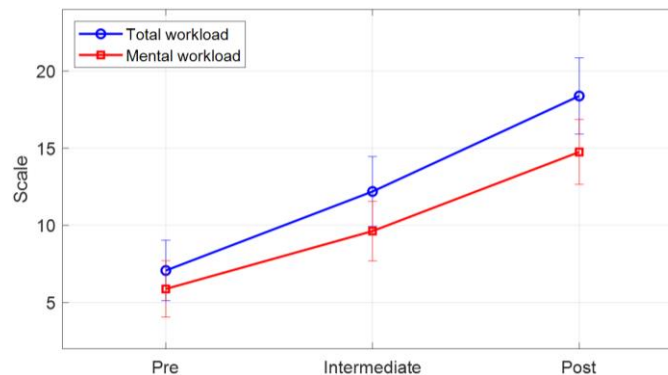
$$Total\ workload_j (participant) = Weight_i \times Scale_i \quad j = Participant$$

Table 15. Calculated weights for six dimensions in NASA-TLX

Dimension	Weight	Dimension	Weight
Mental demand	0.984	Performance	0.0121
Physical demand	0.227	Effort	0.2444
Temporal demand	0.007	Frustration	0.0868

The results of calculating the total workload at the pre-, intermediate- and post-experiment stages using the weights are presented in Figure 47 below. Both total workload and mental workload exhibit an apparent linear increase over time. This trend indicates that the perceived workload, both total and mental, either increases as time progresses or varies consistently depending on a task length. These findings align with numerous studies in the review article [46] showing a positive correlation between workload and participant fatigue, suggesting that fatigue levels may rise as tasks continue. The superior fit demonstrated by the Bagging Trees and Gaussian Process Regression models further confirms that both total and mental workload can be effectively modelled using time as a predictive variable. These regression models demonstrate their capability to predict workload variations over time accurately. Therefore, utilizing time as a predictive variable in regression models plays a crucial role in understanding and forecasting the evolution of mental and total workload as time progresses. This finding underscores the value of incorporating temporal elements in workload prediction models, facilitating better planning and management of workload in various operational settings.

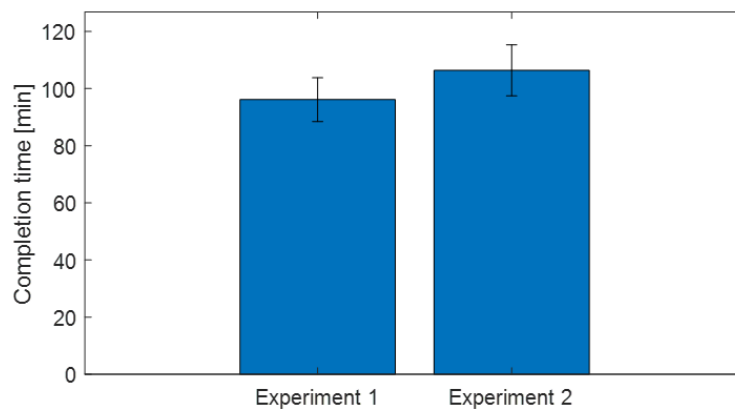
Figure 47. Total workload and mental workload for all participants



6.4.2 Aggregation of NASA-TLX Results from Two Different Experiments

Since each participant completed the experiment at different times, and these completion times varied depending on the experiment type, Figure 48 compares the completion times across different experiments. The completion times include breaks between sessions and the calibration times for the eye-tracking system when required. However, preparation times for data acquisition systems such as EEG, eye-tracking and wrist-wearable sensors before the first session are excluded. The difference in average completion time between the two experiments is approximately 10 minutes with standard deviations of 7.7 and 8.9 minutes, respectively. The study durations between these experiments did not show significant differences. Although the maximum difference in completion time between the two experiments is about 26.6 minutes, the average difference is only around 10 minutes with the standard deviation representing just about 8% of the overall average completion time.

Figure 48. Comparison of experiment completion times in two experiments.



- Experiment 1: Error detection on various visual target types (same lab conditions as the experiment 2)
- Experiment 2: Error/Potential risk detection and user preference study in a driving simulator.

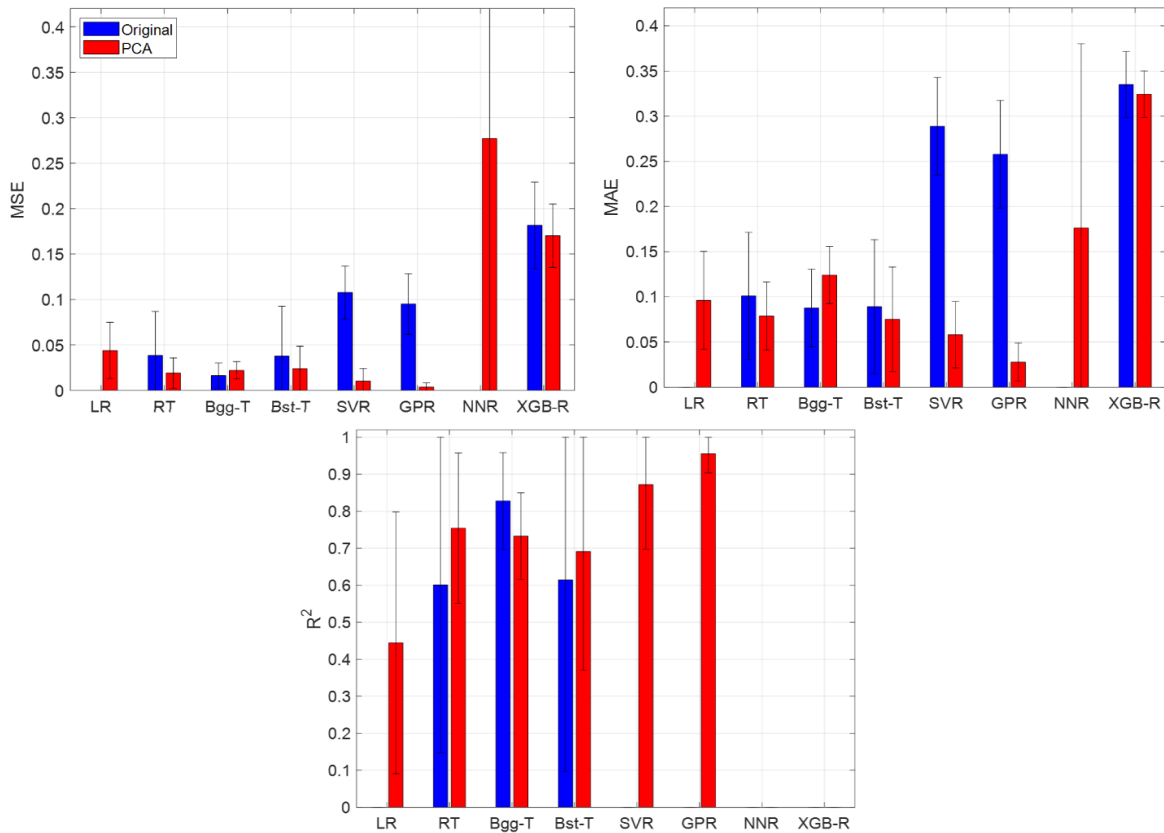
Consequently, the proceeding time used in the regression model was normalized to ensure that variations in completion time did not affect the prediction variable. Additionally, when analyzing the NASA-TLX responses collected to assess participants' workload, the differences in completion time were not considered when comparing responses measured at the pre-, intermediate- and post-experiment stages. The decision to combine the data was based on several considerations that ensured consistency and relevance for the analysis. First, both experiments were conducted under identical lab conditions, using the same monitor and setup to minimize environmental variability. Additionally, the average durations were comparable, with only around 10-minute difference between the two experiments (the Carla simulator and the traffic video review). Second, the primary tasks in both experiments involved watching a monitor without any physical

activity. This ensured that the workload measured was predominantly cognitive, allowing for meaningful comparisons between the two datasets. Third, the workload estimation models in this study were developed using features extracted from physiological signals based on a personalization approach, meaning that each participant had an individualized model. This approach mitigates variability between experiments, as the models were tailored to each individual's data, regardless of differences in the participant groups used in the two experiments. Lastly, the primary purpose of this study was not to compare NASA-TLX scores directly between experiments. Instead, the NASA-TLX data served to identify general workload trends, which were then used to classify workload into four distinct levels. This study focuses on validating the methodology for dividing workload into these classes rather than on specific differences between the experiments. These justifications demonstrate that combining NASA-TLX data was methodologically sound and appropriate for the objectives of the study. However, concerns about combining NASA-TLX results from two different experiments remain. To address this, the aggregation of NASA-TLX results was avoided, and instead, the NASA-TLX data were analyzed independently for each experiment.

6.4.3 Supervised Learning Result

As introduced in Section 4.3.1.1, experiment proceeding time was used as a prediction variable modelled by regression models. To evaluate the performance of the models, we selected Mean Squared Error (MSE), Mean Absolute Error (MAE) and the Coefficient of Determination (R^2) to identify the most appropriate regression models. The fitting of the regression models was evaluated by averaging the performance of personalized models for each participant. Additionally, we compared the results of regression models using all collected features with those using reduced features obtained through Principal Component Analysis (PCA). When using the original features, the 'Regression Tree,' 'Bagging Tree,' and 'Boosting Tree' models were trained stably and demonstrated relatively higher performance compared to other models. On the other hand, after applying PCA to reduce feature dimensions, the 'Gaussian Process Regression' and 'Support Vector Regression' models exhibited higher performance. This indicates that the optimal regression algorithm can vary during the optimization process depending on feature dimension reduction. As shown in Figure 49, the results of the regression models indicate that the 'Bagging Tree' algorithm performed best with the original features, while the 'Gaussian Process Regression' algorithm excelled with the reduced features. The R^2 metric showed an average value of approximately 0.83 for the original features and 0.95 for the reduced features. The highest R^2 values were around 0.95 and 1 for the respective feature sets, indicating that the models nearly achieved maximum performance. Additionally, the other two metrics, MSE and MAE, also demonstrated the best performance with these algorithms.

Figure 49. Performance of the selected regression models (MSE, MAE and R^2)

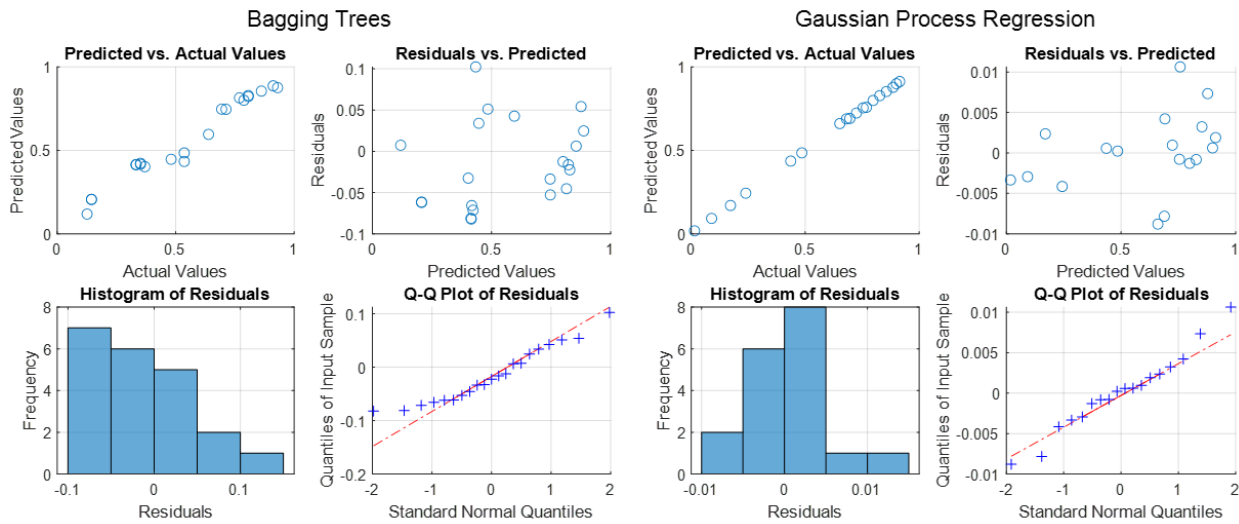


- Eight regression algorithms are tested as follows: Linear Regression (LR), Regression Tree (RT), Bagging Tree (Bgg-T), Boosting Tree (Bst-T), Support Vector Regression (SVR), Gaussian Process Regression (GPR), Neural Network Regression (NNR) and XGBoost Regression (XGB-R).

Furthermore, Figure 50 presents an analysis of the prediction performance for a representative case using the two selected regression algorithms. For the Bagging Trees model with the original features, the Predicted vs. Actual Values plot indicates a relatively good fit with most predicted values closely aligning with the actual values, though some dispersion is noticeable, particularly at higher values. The residuals vs. predicted plot shows a relatively uniform scatter around zero, suggesting an absence of apparent patterns in the residuals and a positive indicator of model performance. The Histogram of Residuals appears approximately symmetric and centred around zero, implying that the residuals are normally distributed. This is further confirmed by the Q-Q Plot of Residuals, where the residuals closely align with the reference line, indicating that they reasonably follow a normal distribution. For the Gaussian Process Regression model, the Predicted vs. Actual Values plot demonstrates a very close alignment between the predicted and actual values, indicating a strong fit with minimal deviation. The Residuals vs. Predicted plot displays

residuals tightly clustered around zero without noticeable patterns, and it suggests an excellent model fit. Although the Histogram of Residuals shows a slight skew, it remains relatively centred around zero and indicates that the residuals are close to being normally distributed. The Q-Q Plot of Residuals further supports this, showing that the residuals align well with the reference line and suggesting a normal distribution of residuals.

Figure 50. Prediction analysis for Bagging Trees for original features and Gaussian Process Regression for reduced features by PCA



These results suggest that if a linear relationship between the time variable and mental workload is established, the time variable could serve as a key factor in predicting changes in mental workload. This finding aligns with the earlier observations from the NASA-TLX results that showed an increase in both mental workload and total workload over time.

6.4.4 Unsupervised Learning Result

For unsupervised learning, we tested multiple algorithms to cluster the feature data: K-means, Hierarchical Clustering, Density-Based Spatial Clustering of Applications with Noise (DBSCAN), Gaussian Mixture Model (GMM), t-distributed Stochastic Neighbor Embedding (t-SNE) and Self-Organizing Maps (SOM). Among these, the K-means and Hierarchical Clustering algorithms produced the most promising results. The clusters formed by these algorithms can potentially represent various levels of mental workload or fatigue states. However, as unsupervised learning does not involve predefined classification or regression, no direct prediction variable represents mental workload or fatigue. Moreover, if the clusters could be mapped to specific class data, they could serve as variables to predict a participant's workload or fatigue

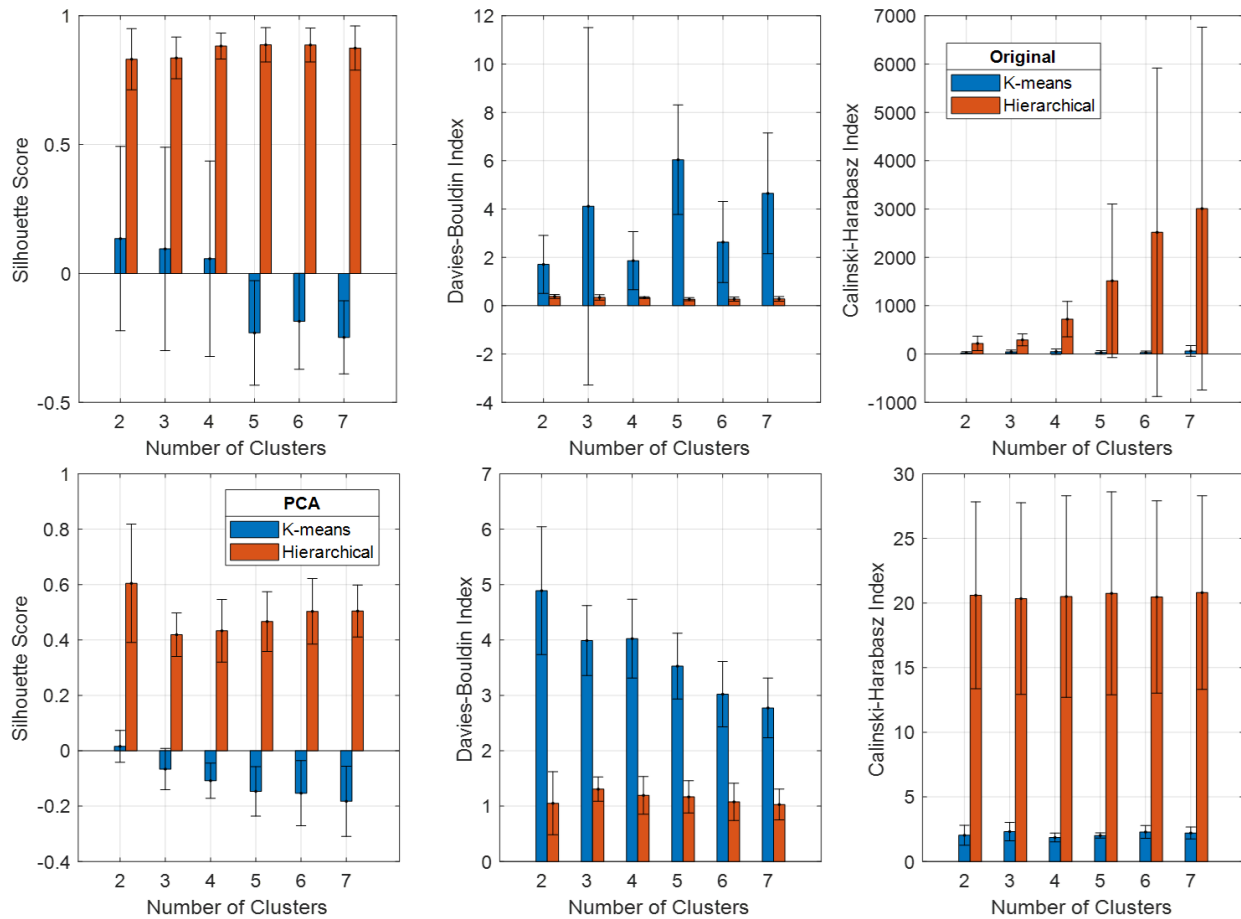
level. However, it remains unclear how each cluster corresponds to specific fatigue states. For example, if there are five clusters, a way to associate each cluster with a mental workload scale from 1 to 5 is needed. We attempted to identify the optimal number of clusters by comparing clustering performance across different cluster counts. To evaluate clustering performance, we used three metrics: Silhouette Score, Davies-Bouldin Index and Calinski-Harabasz Index. Additionally, as with supervised learning, we conducted a comparative performance analysis of the original feature dataset and the reduced feature dataset obtained through PCA. Figure 51 presents a comparative analysis of clustering evaluations using K-means and Hierarchical Clustering algorithms on two datasets: the original feature dataset and a PCA-reduced feature dataset. The clustering quality was assessed across a range of cluster numbers, from 2 to 7, using the three metrics mentioned above. Each metric provides unique insights into the clustering performance for a comprehensive evaluation.

Hierarchical Clustering

The Hierarchical clustering algorithm consistently demonstrates superior performance across all metrics compared to K-means, particularly in the original dataset. The Silhouette Score indicates that Hierarchical clustering is outperformed compared to K-means, often close to or exceeding 0.8, indicating well-defined clusters with good cohesion and separation. This score measures how similar an object is to its cluster compared to other clusters. Scores near 1 suggest that data points are well-clustered and far from neighbouring clusters, while scores near 0 indicate points near the boundary between clusters. The high Silhouette Scores for Hierarchical clustering indicate that data points are well-clustered with clear separation from other clusters. The Davies-Bouldin Index (DBI) for Hierarchical clustering remains relatively low, suggesting high intra-cluster similarity and distinct inter-cluster differences. DBI focuses on both the similarity within clusters and the differences between clusters with lower values indicating better clustering. The low DBI values across different cluster numbers highlight Hierarchical clustering's ability to form cohesive and well-separated clusters. The Calinski-Harabasz Index (CHI) for Hierarchical clustering is moderate but consistent, further supporting the effectiveness of the algorithm. CHI evaluates the ratio of the sum of between-cluster dispersion to within-cluster dispersion with higher values indicating better-defined clusters. The consistent CHI values for Hierarchical clustering suggest that the clusters are well-defined and balanced in terms of dispersion. When considering the number of clusters, Hierarchical clustering generally shows that two clusters provide the best results across all metrics. The Silhouette Score remains high and stable across different cluster numbers, which indicates that clustering performance does not significantly degrade with varying cluster counts. This stability suggests that Hierarchical clustering can effectively manage different cluster configurations, but optimal performance is achieved with fewer clusters, simplifying feature classification. When considering the trade-off between the number of clusters

and clustering performance as well as the standard deviations of metrics for each cluster count, it becomes evident that while two clusters perform well, superior clustering performance is achieved with four clusters. In the PCA-reduced dataset, Hierarchical clustering continues to outperform K-means, though with generally lower scores compared to the original dataset. The reduction in the Silhouette Score and the increase in the Davies-Bouldin Index suggest that PCA introduces some level of ambiguity, affecting cluster quality. Nevertheless, Hierarchical clustering maintains relatively better performance than K-means, highlighting its robustness.

Figure 51. Comparison of clustering performance results in various cluster numbers (Original features vs Reduced features by PCA)



K-means Clustering

K-means clustering, while effective, does not perform as well as Hierarchical clustering. In the original dataset, the Silhouette Score for K-means is positive but lower than that of Hierarchical clustering, which indicates that the clusters are reasonably defined but exhibit less cohesion and separation. The DBI

increases with the number of clusters. It suggests that as more clusters are introduced, the quality of clustering decreases due to higher intra-cluster similarity and less distinct inter-cluster differences. However, the CHI shows higher values with an increasing number of clusters, indicating better-defined clusters when more clusters are used. In the PCA-reduced dataset, the performance of K-means clustering declines further with lower Silhouette Scores and higher DBI values. This reduction highlights the impact of PCA on clustering quality, introducing more ambiguity and diminishing the overall effectiveness of K-means clustering. Despite this decline, K-means shows a notable increase in CHI values as the number of clusters increases, and it suggests that the between-cluster dispersion improves with a greater number of clusters while intra-cluster similarity may suffer.

6.4.4.1 Optimal Cluster Number

Across both algorithms and datasets, there appear to be two best-performing clusters. This is evidenced by the consistently high Silhouette Scores and low DBI and CHI values for two clusters, which suggest that fewer clusters provide a more straightforward and effective classification of features. However, the four clusters are also performing well, showing good results across all metrics. The stability of clustering performance with two and four clusters indicates that this configuration simplifies the classification process and distributes clusters for multiple stages in either total workload or mental workload, optimizing the complexity and enhancing the interpretability of the results.

6.4.4.2 Discussion

Without a feature reduction process, Hierarchical clustering with two and four clusters demonstrates strong clustering quality. The performance remains relatively stable across different cluster numbers, indicating that the algorithm can manage varying configurations without significant degradation. The slightly lower performance observed with two clusters compared to more clusters may be due to the closer proximity of features within the same cluster, which reduces perceived separation. In the PCA-reduced dataset, both algorithms experience a decline in clustering performance. However, hierarchical clustering maintains a better overall performance than K-means. Although PCA reduction introduces more ambiguity and affects clustering quality, hierarchical clustering still manages to produce relatively well-defined clusters.

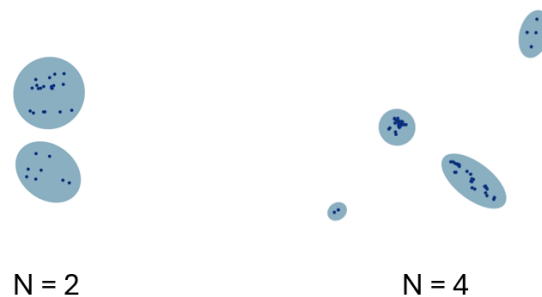
In the next chapter, based on the clustering performance of unsupervised learning algorithms and the results regarding the appropriate number of clusters, we will attempt to convert the clustering results of each feature data row into labelled data using the NASA-TLX data from pre-, intermediate- and post-experiment stages. Using the converted labels, we plan to develop a model for predicting participants' workload through label

propagation in semi-supervised learning. The results will then be compared with the models built in Section 4.5.1 on the regression model that used experiment proceeding time as the prediction variable.

6.4.5 Semi-Supervised Learning Result

Based on the previous results in unsupervised learning (clustering algorithms), we decided to focus on binary classification and four classes for workload that showed the best clustering results based on reduced dimension data (2D). Clustering examples are shown below (Figure 52).

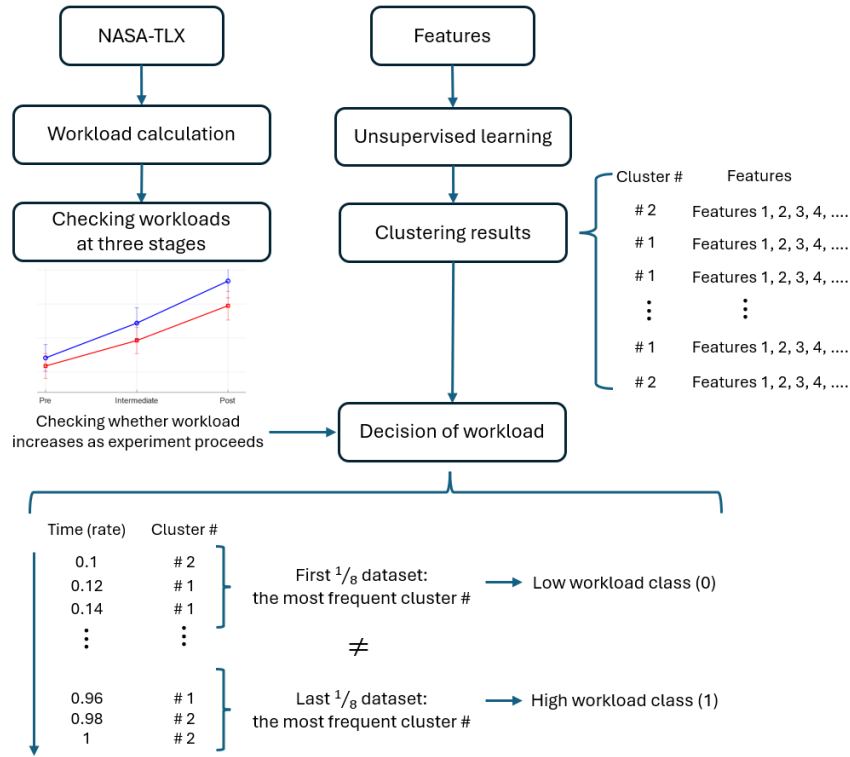
Figure 52. Clustering examples based on the number of clusters (2 and 4 clusters)



6.4.5.1 Condition 1: Binary Classification of Workloads (2 Clusters)

In Figure 53, there is a class decision process of binary workload classes. Each feature data row was classified into one of two workload classes based on a clustering result. The two identified clusters were subsequently mapped to two distinct classes. In this process, we first compared the most frequently occurring cluster in the first 1/8 of the dataset with the most frequent cluster in the last 1/8 of the dataset. In the clustering results, we compared the frequency of the two clusters within the first 1/8 of the dataset. The most frequent cluster appeared approximately 60-90% more often than the other cluster. Additionally, if these two clusters differed, they were classified as representing low and high workload classes, respectively. This result aligned with the expected increase in workload over time, which was similarly observed in the NASA-TLX results. For all participants except two participants, the cluster data aligns well with the workload measured in the NASA-TLX by accurately categorizing the two clusters into low and high workload classes. Based on these results, we performed label propagation, which assigns labels for the low workload class (1) and high workload class (2) to each cluster.

Figure 53. Decision process of binary workload classes

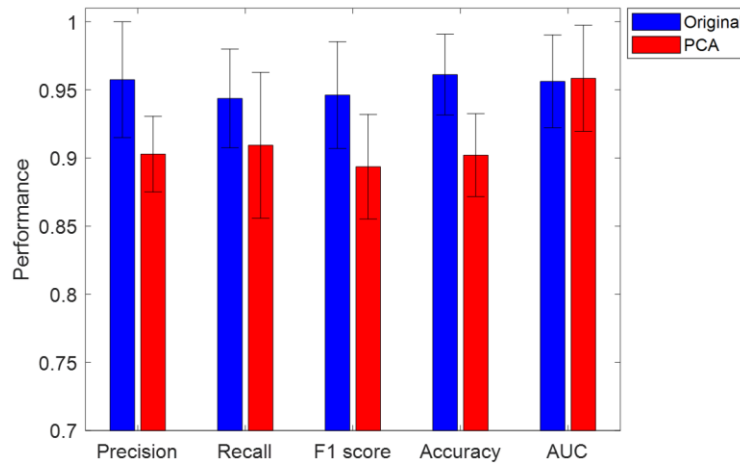


For the two participants whose workload classes did not match with the NASA-TLX data, the unsupervised learning-based clustering revealed the following characteristics: the feature data measured during the intermediate stage of the experiment were classified into the high workload class (2) while the feature data measured during the initial and final stages were classified into the low workload class (1). Using the dataset with defined classes for binary workload levels, we proceeded to perform workload classification modelling. Therefore, we are able to build a classification model based on the workload classes matched to each cluster by considering cluster distances and rankings. Figure 54 presents the performance of the classification model. It showed all metrics over 95% performances with original features and 90% performances with reduced features by PCA. It showed fairly good performances regardless of feature reduction.

6.4.5.2 Condition 2: Multi-class Classification of Workloads (4 Clusters)

Similar to the approach with two clusters, we design a decision process of four workload levels based on clustering results, as shown in Figure 55. The clusters with varied distances by unsupervised learning are illustrated in Figure 56. The algorithm assigns workload levels to identified and unidentified clusters based on their proximity and temporal distribution across the dataset.

Figure 54. Classification performance based on matched workload classes with clusters



In the first step, binary workload levels are initially established to classify the first 1/8 (beginning), intermediate 1/8 (middle), and last 1/8 (end) segments of the dataset, reflecting variations in workload over time. In the next step, the most frequent clusters are identified within these segments, representing distinct workload levels. The following step assigns specific workload levels to the identified clusters based on NASA-TLX measurements, categorizing them as the lowest (Level 1), highest (Level 4), or intermediate levels (Levels 2 and 3). For unidentified and undecided clusters, the final step defines their workload levels based on proximity to the identified clusters. For example, an unidentified cluster positioned closer to a low workload cluster is assigned as an intermediate level (e.g., Level 2). In Figure 56, when workload increases toward the highest level at the end of the session, Cluster 1 represents the initial workload level (Level 1), Cluster 4 corresponds to the highest workload level (Level 4), and intermediate clusters (e.g., Cluster 3) are assigned Levels 2 or 3 based on their proximity to the other clusters. This approach ensures a structured and dynamic classification of workload levels across the dataset, accommodating both temporal patterns and spatial relationships among clusters. However, for the two remaining participants, the clustering results from each 1/8 segment revealed clusters with overlapping or ambiguous highest frequencies, making it difficult to correctly define 1-2 clusters. In such cases, we ranked the clusters based on the distance results obtained from the clustering algorithms. Combined with the already defined clusters, this ranking allowed us to determine the overall cluster ranks and associate them with the corresponding workload levels. Figure 57 presents the performance of the classification model. It showed all metrics over 95% performances with dimension-reduced features by PCA and 90% performances with original features. It showed great performances regardless of feature reduction, but the dimension-reduced features performed better in the higher granularity model of workload estimation.

Figure 55. Decision process of multi-workload classes (Four workload levels)

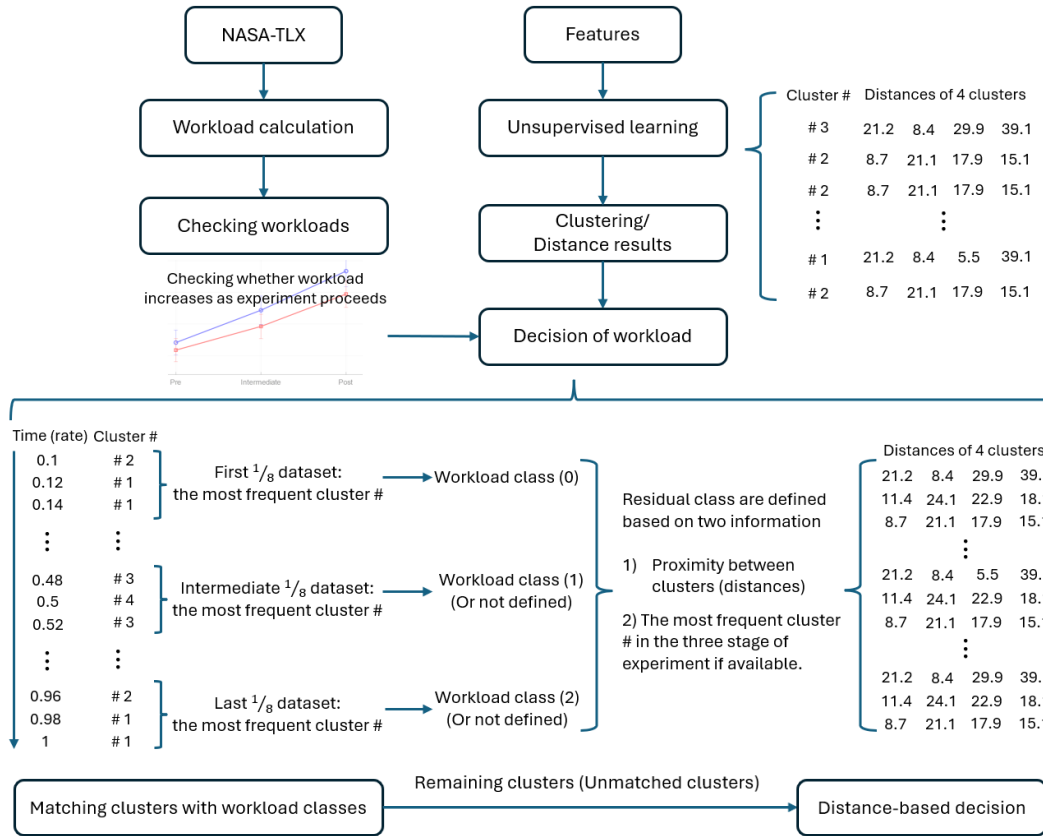
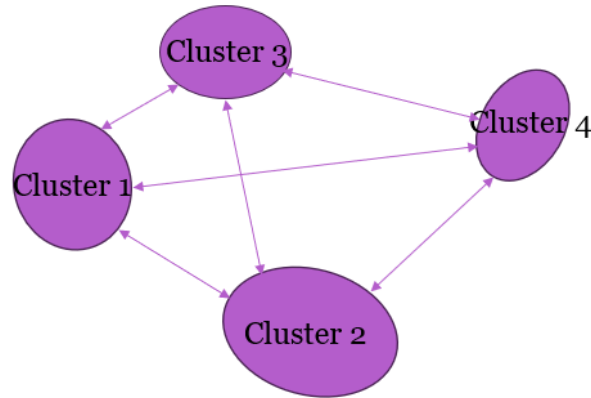


Figure 56. Illustration of Clusters: Mapping Clusters to Workload Classes for Four workload levels



As a result of the decision-making processes for both binary and four-class workload models, Figures 58 present the percentage distribution of workload levels under these two conditions: binary classification and four workload classes (representing varying severity levels of workload). Both figures indicate that as the experiment duration increases, the estimated workload levels rise, demonstrating a clear trend of increasing

workload over time. Specifically, in the final phase of the experiment (80-100% of the total duration), the highest workload levels show a marked increase in both binary and four-class models. Although some instances of high workload levels appear in the early stages of the experiment, these occurrences can be interpreted as participant-specific variations in workload fluctuation timing, as well as in the extent to which estimated workloads rise and fall over the course of the experiment. Nevertheless, statistical analysis across all participant data reveals a distinct difference in the frequency of estimated workload levels across the early, middle, and late stages of the experiment. This finding is consistent with the results obtained from the earlier NASA-TLX analysis.

Figure 57. Classification performance based on matched workload classes with clusters

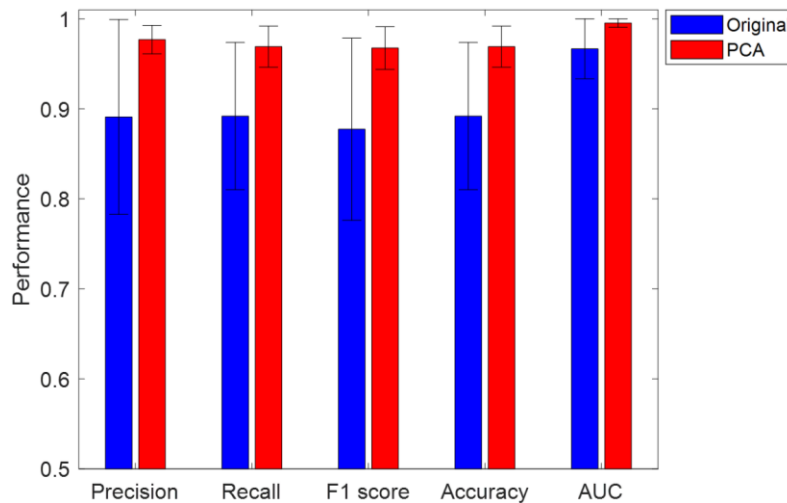
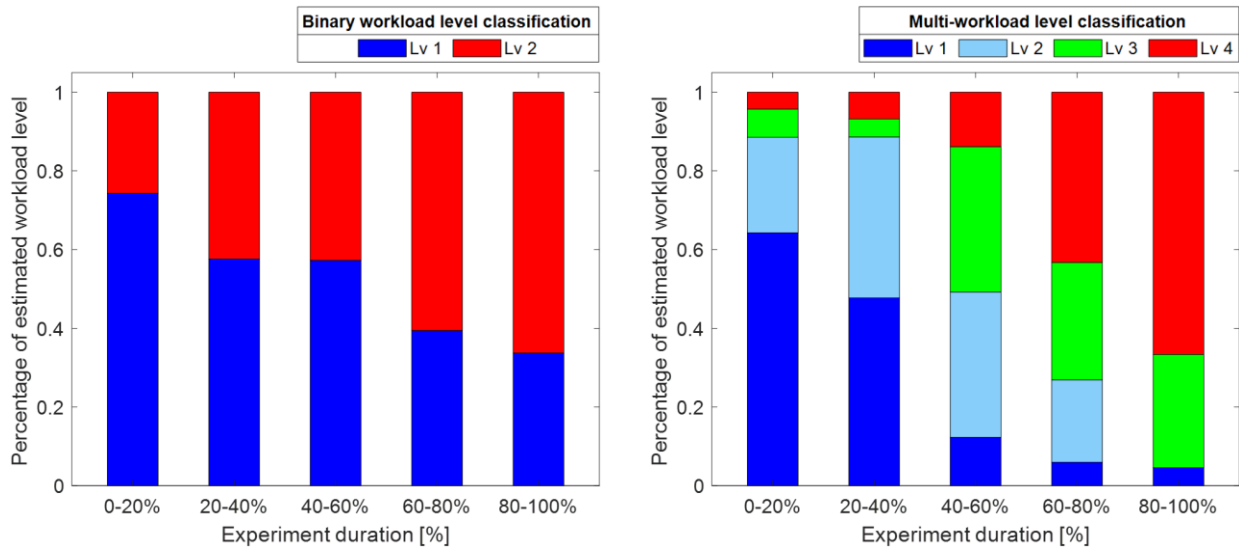


Figure 58. Percentage of binary workload levels (2 Levels) and multiple workload levels (4 levels)



6.5 Feature Analysis of Workload Estimation Models

This section provides an in-depth analysis of the importance of features in classification models for binary and multi-class workload classification, as shown in Figure 59. The feature dataset includes features extracted from EEG signals (Cz, Fz, Pz), pupil size (PS), pupil center degree changes (PCD), pupil center magnitude changes (PCM), body temperature (Tem), electrical dermal activity (EDA) and heart rates (BVP + HR). The importance of these features was evaluated using ANOVA, Chi-squared test, Mutual Information, Lasso, Permutation and BorutaSHAP or SHAP methods. The left side in Figure 59 represents the feature analysis results in the binary classification (two classes of workload levels), while the right side corresponds to the results in the multi-class classification (four classes of workload levels).

6.5.1 Binary Workload Classification

In the binary classification scenario, features extracted from EEG signals demonstrate the highest importance, particularly in the Pz channel. This suggests that the Pz channel in EEG recordings is highly influential in distinguishing between the two classes. Two pupil features (PS and PCD) also exhibit significant importance, indicating their relevance in the binary classification model. Conversely, features related to EDA and BVP + HR show minimal importance across all evaluation methods, suggesting that these physiological signals may not be as critical in distinguishing between the two classes in this context. The ANOVA method highlights the Pz channel features as the most important feature, followed by PS and PCD, while EDA and BVP + HR features remain the least important. This trend is consistent with the results from the Chi-squared test, where Pz and PS features are significant, and EDA features show minimal importance. Mutual Information also emphasizes the significance of Pz and PS features, although it assigns lower importance to EEG channel features compared to ANOVA and Chi-squared. BorutaSHAP results align with these findings, showing high importance for features from the Pz channel and PS and minimal importance for EDA features. Lasso further reinforces this trend by emphasizing PS and Pz features while minimizing the importance of EDA features. The Permutation method also indicates that PS and Pz features dominate in importance, with EDA and Tem features being less significant.

6.5.2 Multi-class Workload Classification (4 Classes ~ 4 clusters)

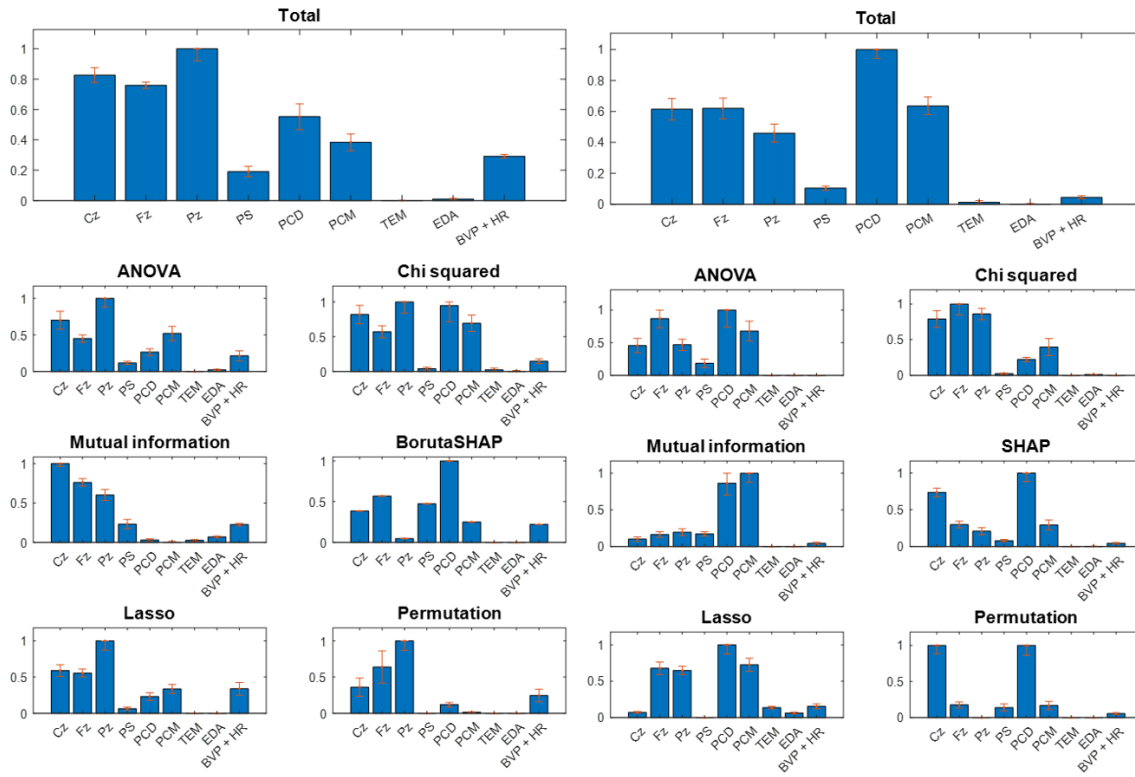
In multi-class classification, the importance of features differs compared to binary classification. The PCD feature emerges as the most important, followed by the PCM feature, which is comparable to the Cz and Fz features in the multi-class workload model. In the binary workload model, the Pz feature is the most important, followed by other EEG features. While the rankings of pupil features remain the same in both binary and multi-class workload models, the scale of their importance shows a significant difference. The

most notable shift in feature importance between the two workload models is from Pz to PCD, suggesting that as the number of workload classes increases (indicating more complex classification tasks), the importance of features for distinguishing classes shifts from being centred on neural activities to focusing on pupil-related features. Additionally, the importance of the BVP + HR feature significantly decreases, reaching nearly zero as the number of workload classes increases. Excluding the Chi-squared test, all other feature importance measurements indicate that the PCD feature is vital in the multi-class workload model. However, in the binary workload model, the Pz and PCD features show relatively higher importance than other features in some methods, but this pattern is not consistent when compared to the multi-class workload model. Other physiological features, such as those related to electrodermal activity (EDA) and body temperature (Tem), consistently show minimal importance across all methods and models.

6.6 Adaptive Models of Workload Estimation

Figure 9 indicates that some participants exhibit stable results even in higher cluster scenarios. This suggests that it is possible to adjust the number of classes within the workload classification model. Furthermore, as shown in Figures 12 and 14, the classification performance does not drastically decline when the number of classes is adjusted (e.g., from binary to four workload classes). This flexibility allows for personalization when building an individual's workload prediction model; by adjusting the desired number of classes, one can refine the predictive workload levels according to specific needs. This adaptability is beneficial for creating a model that can be used in various situations by adjusting the number of classes. For example, if a driver perceives their initial workload level to be high, they can set an appropriate number of desired classes. This allows the system to respond effectively if a workload level higher than the initial one is estimated. Such an approach can serve as a proper alert mechanism before obvious symptoms of fatigue become apparent. This is because the drivers may already be highly fatigued at the time distinctive signs of fatigue are observed, and it is often too late to prevent an accident. Fatigue-related accidents often occur because drivers fail to recognize the early signs of drowsiness, such as micro-sleeps or nodding off [275]. By the time these signs become noticeable, drivers have often already reached a state of severe impairment where their ability to react is significantly diminished. Therefore, during the transition to intermediate or high workload levels, the adaptive prediction system can appropriately recommend breaks or suggest other methods of refreshment to the driver.

Figure 59. Feature importance results in binary and multiple classifications: Two workload classes (Left) and Four workload classes (Right)



6.7 Discussion

The findings presented in this chapter illustrate the effectiveness of using a combination of supervised, unsupervised and semi-supervised learning methods to develop predictive models of mental workload based on physiological signals and eye-tracking data. The study highlights the importance of selecting appropriate features and clustering techniques to accurately classify workload levels, particularly when utilizing NASA-TLX results as a reference for model development. In the supervised learning context, the models demonstrated strong predictive performance, mainly when using experiment proceeding time as a prediction variable. The Bagging Tree and Gaussian Process Regression models emerged as the top performers with high R^2 values and low error metrics (MSE and MAE). This suggests that these models are well-suited for predicting mental workload over time. The exploration of unsupervised learning methods, specifically K-means and Hierarchical Clustering, provided valuable insights into how clustering can reveal underlying patterns in workload states. Hierarchical Clustering consistently outperformed K-means, especially in the original feature dataset, suggesting its robustness in handling complex data structures. The study further identified that fewer clusters (e.g., two or four) tended to provide more stable and interpretable results. It also simplifies the classification process. Moreover, the application of semi-supervised learning

allowed for the integration of both labelled and unlabeled data, and it can enhance the model's ability to predict workload levels in scenarios where labelled data is scarce. This approach successfully classified feature data into workload classes, demonstrating the utility of combining clustering results with NASA-TLX data to refine workload predictions. The feature importance analysis across binary classification models highlighted the critical role of EEG signals, particularly the Pz channel, in distinguishing between different workload levels, but multi-class classification models show similar feature importance across all three EEG channel features. PCD and PCM also emerged as significant contributors more than pupil size features, especially in more complex multi-class classification tasks.

6.8 Limitations

In this study, I proposed an approach to estimate cognitive workload levels prior to being highly fatigued and to customize the granularity of estimated workload classes (personalization), such as from binary to multi-class levels of 3 or more. However, the effectiveness of this approach has not yet been extensively validated. Below, several limitations are outlined. First, the estimated sub-levels of cognitive workload (between mild and moderate/high prior to a severe stage) might not be linearly correlated, while quantitative measures and self-reported workload may align [272, 273]. In other words, it may not guarantee that the sub-level divisions reflect accurate workload distinctions. Second, when the workload levels are divided into too many finely tuned segments, especially between mild and moderate workloads, it results in an increase in the number of clusters representing the complexity of estimation. This can create ambiguity between clusters (workload levels). For instance, minor differences in neighbouring levels may not be meaningful or detectable with the same clarity. Third, during the study, participants kept engaging in tasks, and none exhibited clear signs of severe cognitive workload, such as significant changes in eye-blinking rate, saccadic movements, dozing off, heart rates or specific EEG patterns in the brain, etc [274]. Fourth, establishing accurate ground truth values for cognitive workload levels presents a significant challenge. In this study, we classified these levels based on individual participants' NASA-TLX ratings, defining classes through personal assessments. This personalized modelling approach lends credibility to the ratings as ground truth values. However, objective indicators, particularly in pupil dynamics, such as saccade velocity, blink rate, and eyelid closure rate, did not exhibit notable patterns in this study. These features often emerge in states of severe fatigue and can serve as objective measures for ground truth values. The absence of such objective measures in our study highlights a gap. Moreover, limited research exists on estimating cognitive workload levels through significant changes in pupil dynamics or other physiological signals before reaching severe states. Detecting severe states via pronounced physiological changes is relatively straightforward; however, such detection may occur too late to prevent risky situations. Therefore, it's

imperative to explore objective measures, like significant changes in pupil dynamics or physiological signals, for intermediate cognitive workload levels. Additionally, when adapting the granularity of workload levels, estimating the workload of unidentified clusters between known levels may be inaccurate, as the progression may not be linear. Nonetheless, physiological and eye-tracking signals can significantly impact human state estimation research. Even if the progression isn't linear, clear distinctions between clusters can emerge if instances are sufficiently differentiated from identified clusters. Despite this, accurately estimating the workload levels of unidentified clusters necessitates more precise ground truth values. Addressing these ground truth issues may require additional experiments and more frequent assessments, such as repeated NASA-TLX evaluations, to mitigate ground truth ambiguities.

These signs are typically associated with high cognitive workload. To address the limitations above, we will explore the following topics:

- 1) Quantitative analysis for correlation between self-assessment results and cognitive tests.
- 2) Inclusion of clusters for high cognitive workload states.
- 3) Investigation of various factors' impacts on cognitive workload: initial driver states, driving mileages in a day, traffic volume, road type, traffic speed, driving performance, body movements, changes in gazing patterns with salient events/objects while driving, user manual input for a sensitivity level to detect changes in cognitive workload.
- 4) Extensive research for optimizing cluster segmentation for various factors above.
- 5) Investigation of our approaches for comprehensive workload by all dimensions in the NASA-TLX, e.g., physical, temporal demand, task performance, etc.
- 6) Enhance the reliability of defined workload levels (including levels that defined by cluster distances) by conducting frequent assessments using NASA-TLX or other measurement tools.

We believe that addressing these limitations can greatly advance our methods for developing accurate driver cognitive workload estimation models and extend their use to practical applications in advanced driver assistance systems. This progress will support the design of systems that are better aligned with drivers' cognitive demands, ultimately enhancing safety and driving efficiency.

6.9 Conclusion

This study successfully addressed research questions 6~9 (RQ 6~9) and demonstrated the potential of ML techniques in estimating mental workload using physiological signals. Our approach addresses some issues:

1) limited workload class data and 2) intermediate stage of cognitive workload/fatigue that may not show any symptoms at severe stages (RQ6). Using a combination of supervised, unsupervised, and semi-supervised learning methods, the multi-class workload estimation models developed in this chapter demonstrate the capability to adjust the granularity levels of workload classes (RQ7). For RQ8, while the model performance did not decline with increased granularity levels (~ increased number of classes and model complexity), the cases currently available lack sufficient results across a wider range of granularity levels. Thus, RQ8 is only partially addressed in this study. To fully address this research question, future studies should investigate additional granularity levels to provide a more comprehensive understanding of their effects on model performance. In the adjustable model approach, users can control the granularity of the workload levels, which is connected to a driver monitoring/ warning system activated when a certain workload level is estimated. Furthermore, the findings underscore the importance of feature selection and clustering techniques in enhancing model performance (RQ9). Our study contributes to the field of cognitive workload estimation. We provide a comprehensive approach to customizable cognitive workload estimation models by leveraging the three learning concepts. It can be applied to a user-centred driver monitoring/warning system for driver safety improvement.

Chapter 7. Perception Model of User Preference for Self-Driving Behaviours

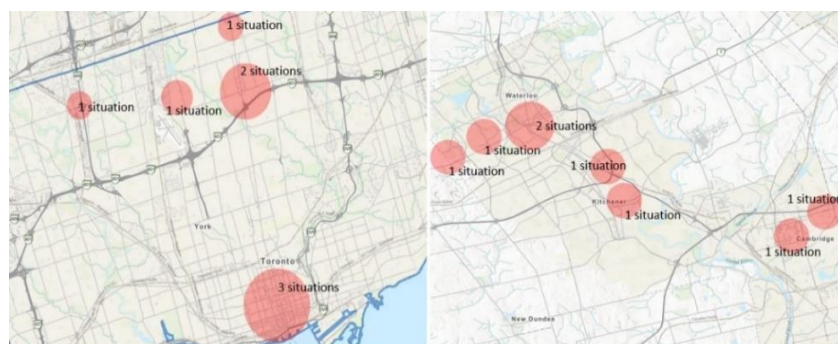
In this chapter, our primary goal is to develop perception prediction models for user preferences regarding self-driving behaviours. Before constructing these models using physiological recording and eye-tracking systems, it is essential to understand how user preferences change in various traffic situations. To achieve this, we conducted a survey to investigate user preferences for self-driving behaviours and used the data to build predictive models first to see if user preferences are predictable by ML models. After achieving survey analysis and ML model based on survey data, we tried to develop predictive models for user preferences in different traffic situations and identified various factors influencing these preferences. By verifying whether these factors are reflected in physiological signals and eye-tracking data, we aim to create robust predictive models using these multimodal data.

7.1 Study Design

7.1.1 Selected Traffic Situations

In Figure 60, traffic situations were recorded in some areas. The recorded areas were selected after multi-day hours-long naturalistic driving in these two regions. The recorded videos were divided into a few groups of situations: driving in Toronto, Waterloo, roundabouts, driving highway to city and city to highway. Among the 18 video clips, 16 were recorded in the city of Toronto (Toronto) and the regional municipality of Waterloo (Waterloo), representing two traffic complexities, high and low, respectively. Selection of video clips takes factors such as traffic volumes, locations, potential hazards and required decision-making moments into consideration. The remaining two videos were recorded in high- and low-traffic roundabout situations that were not marked on the map below.

Figure 60. Areas where selected videos were recorded the city of Toronto (Left) and the regional municipality of Waterloo (Right).



Particularly, the hazards of each situation in the two traffic complexities were divided into real (materialized) and potential hazards. Some of the hazards included a right turn blocked by a police vehicle, parked vehicles and pedestrians, a vehicle with emergency lights and a sudden lane change. In the two traffic complexities, hazard types could not be exactly matched since video recording proceeded under natural driving conditions. So, each situation could not be controlled, and minor traffic events were excluded from the main traffic events. In Table 16, situations with high traffic complexity (volumes) were counterbalanced with situations with low traffic complexity (volumes), and screenshots of all 18 situations are shown. In the online survey, video clips were mixed to avoid participants directly comparing each situation in high and low traffic complexities.

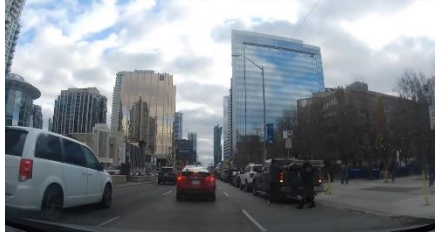
Table 16. Counterbalanced situations and their descriptions

Complexity	High traffic (9)	Low traffic (9)
City scenarios	Toronto (6)	Waterloo (6)
Roundabout	Roundabout high traffic (1)	Roundabout low traffic (1)
Highway	Highway to Toronto (1) Toronto to Highway (1)	Highway to Waterloo (1) Waterloo to Highway (1)

Table 16. (Continued)

Situation	High traffic complex (Toronto) potential hazard: Higher	Low traffic complex (Waterloo) Potential hazard: Lower
S1: Police	 Blocked a right turn by a police vehicle. Speed: < 30 km/h (limit: 40 km/h) Real: a vehicle blocked right turn and lane change due to the narrowed lane. Potential: sudden stop of front vehicles.	 A police vehicle parked with a vehicle pulled over on the right-side. Speed: ~40 km/h (limit: N/A) Real: vehicles stopped on the right lane and slowed down due to the narrowed lane. Potential: N/A

S2:
Multi-
situation



Lots of parked vehicles and pedestrians on the right side.
Speed: ~40 km/h (limit: 40 km/h)
Real: vehicles parked, pedestrians near the cars on the right.
Potential: chance of an unexpected merging of parked vehicles with flashing



Two parked vehicles on the right side and a vehicle coming on the other side.
Speed: ~40 km/h (limit: 40 km/h)
Real: slowed down due to vehicles parked in the right lane, narrowed lane, and a vehicle passing on the other side.
Potential: N/A

S3:
Driving
around
car



In relatively busy traffic, a vehicle with an emergency signal after a front vehicle is away.
Speed: ~50 km/h (limit: N/A)
Real: a vehicle with emergency lights and sudden lane change
Potential: the vehicle's movement.



Changing lanes behind the bus slowing down at the bus stop.
Speed: ~50 km/h (limit: N/A)
Real: stopped bus and limited sight
Potential: vehicles coming out from the exit next to the bus stop.

S4:
Parking
lot



Lots of vehicles, and pedestrians coming out (busy traffic).
Speed: ~20 km/h (limit: N/A)
Real: a vehicle exiting a parking spot, pedestrians in the middle of the parking lot.
Potential: N/A



Only three pedestrians walking at a safe distance.
Speed: ~20 km/h (limit: N/A)
Real: Slowed down due to pedestrians walking around.
Potential: N/A

S5:
Turn
Right



People crossed and vehicles merged into the first lane.
Speed: ~40 km/h (limit: 40 km/h)
Real: pedestrians, cyclists crossing the road and a lane change
Potential: N/A



Turning right after a pedestrian crossing.
Speed: ~40 km/h (limit: N/A)
Real: a pedestrian crossing.
Potential: N/A

S6:
Turn Left



Busy traffic but stopped at a signal. A limited sight to check coming vehicles on the merging lane.
Speed: ~40 km/h (limit: 50 km/h)
Real: limited sight and large traffic.
Potential: traffic of merging lane



Turning left at intersection w/o risks in a residential area.
Speed: ~30 km/h (limit: 30 km/h)
Real: stopped at the intersection before turning left.
Potential: N/A

S7:
Roundab
out



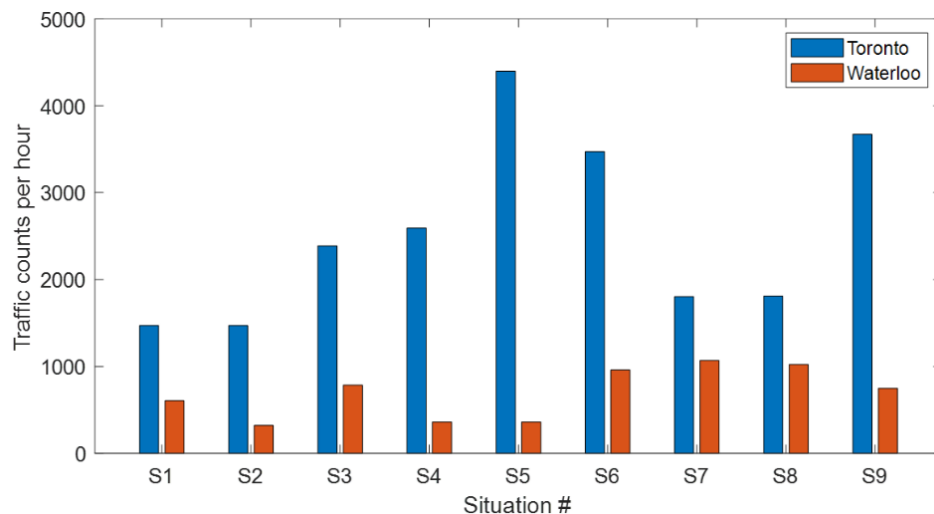
High traffic complexity in a roundabout.
Speed: ~30 km/h (limit: N/A)
Real: busier traffic and a truck.
Potential: vehicles entered from other sides.



Low traffic complexity in a roundabout.
Speed: ~40 km/h (limit: N/A)
Real: N/A
Potential: N/A

S8: Highway to City	 High traffic complexity in an expressway to Toronto downtown. Speed limit: 90 km/h to 50 km/h Real: N/A Potential: busier traffic in Toronto.	 Low traffic complexity in a highway to a region of Waterloo. Speed limit: 100 km/h to 60 km/h Real: N/A Potential: traffic in Waterloo.
S9: City to Highway	 High traffic complexity from a suburban region of Toronto to a highway. Speed limit: 60 km/h to 100 km/h Real: N/A Potential: Merging to highway.	 Low traffic complexity in a suburb region of Waterloo to a highway. Speed limit: 70km/h to 90 km/h Real: N/A Potential: Merging to highway.
	- Real: Materialized hazard	- Potential: Potential hazard

Figure 61. The comparison of the average traffic counts per hour in each situation of the two regions (S#: Situation number).



To provide a comparison of the traffic flows in each situation of the two regions, Figure 59 shows traffic counts per hour data collected from government datasets and reports [275, 276]. Hourly traffic counts of the situations in Toronto were 1.5 times up to 10 times larger than those in Waterloo.

7.1.2 Preference Options for Self-Driving Behaviours

In this study, we delved into two distinct preference classifications: the binary preference of 'prefer' and 'not prefer,' and the preference for multiple actions of AVs designed by the previous studies [277, 278] as shown in Table 17. Beyond a binary preference for self-driving modes, offering users multiple choices for self-driving actions can contribute to a highly user-centred design in autonomous features [279, 280]. Furthermore, the details of such actions can be structured using a two-staged approach incorporating informing and actual actions. This approach recognizes that drivers may only require an 'informing' action from the AV system without necessitating a takeover of control in certain traffic situations. In our exploration of multiple actions, we grouped the multiple actions into different categorizations shown in

Table 17. AV behaviour options designed by Park et al. [277] and Billings [278].

		Act		
		Act autonomously	Receive consent	Don't act
Inform	Inform before	Inform-Act	Inform-Consent	Alert-Handover
	Inform after	Act-Inform	N/A	N/A
	No Inform	Act-No Inform	N/A	N/A

Table 18 to further investigate user preferences for each action categorization. This approach can provide a deeper understanding of user preferences than user binary preferences and aids in tailoring self-driving features to align with their needs and expectations.

Table 18. Categorizations of preferred actions based on different aspects: user preference, informing action, and action agent [242].

Categorizations	Preferences	Informing	Action agent
Act-No Inform	Preferred self-driving	Not being informed	AV-engaged (Human-disengaging)
Act-Inform		Being informed	
Inform-Act			
Inform-Consent	User choice		Human-engaging
Alert-Handover	Preferred takeover		

7.1.3 Participants

7.1.3.1 Survey Participants

We conducted participant recruitment exclusively in the United States and Canada. A total of 175 individuals are from two distinct age groups: 23-44 (young to middle-aged) and 60+ (older than 60). We selected age groups based on the existing survey studies [281, 282]. Liljamo et al. (2018) highlighted a significant disparity in attitudes towards self-driving technology between individuals aged mid-20s ~ mid-30s and mid-50s ~ mid-60s with this divide widening with age [281]. Conversely, Lee and Hess (2020) found that participants in their mid-20s ~ mid-40s exhibited the highest comfort levels with autonomy, contrasting sharply with people in their mid-60s whose comfort levels declined [282]. Consequently, we focused on these two groups, slightly adjusting age ranges to ensure the inclusion of active drivers familiar with autonomy: mid-20s to 23 years for the younger group and mid-60s to 60 years for the Old group, resulting in age brackets of 23-44 and 60+ respectively with a 16-year gap between them. A prerequisite for participation was possession of a valid US or Canadian driver's license with exclusion criteria extending to individuals holding probationary or learner's permits. Following recruitment, five participants displaying low-quality survey responses characterized by excessive repetition (over 90%) were excluded from the analysis. A detailed overview of participant demographics, e.g., gender distribution and age statistics, is shown in Table 19. The number of participants recruited for this study ensures a 0.35 effect size at a power level of 0.95 and an alpha error probability of 0.05, which means the study is highly capable of detecting moderate differences or relationships with a very low probability of Type I and Type II errors. This level of statistical precision ensures that the findings are robust and reliable, providing meaningful insights into the phenomena under investigation.

Table 19. Age and gender information.

Age groups	Y-M: 23-44 years old	Old: 60+ years old
Average (SD)	31.4 ($\pm$ 5.7)	64.7 ($\pm$ 4.5)
Men	29	36
Women	55	49

7.1.3.2 Experiment Participants

For the lab experiment sessions, we recruited twelve participants (8 men and 4 women) with an average age of 29.6 years and a standard deviation of 2.7 years. The number of participants in these sessions is not comparable to our online survey research, which included over 100 participants, as the lab sessions cannot be automated. Additionally, the lab experiments require the constant presence of an investigator to

appropriately handle unexpected situations or any discomfort experienced by the participants. Furthermore, these sessions involve multiple interactions and require participants to wear three different systems: EEG, eye-tracking devices, and wrist-wearable sensors. These logistical challenges make it difficult to recruit a larger number of participants. The recruitment of 12 participants for this study was determined based on prior research demonstrating that within-subject designs with moderate to large effect sizes (around 0.6) require approximately 13 participants to achieve 80% power at an alpha level of 0.05. Existing studies in the fields of user cognition and perception, which are closely related to user preferences and involve physiological and eye-tracking signals, recruited between 5 and 21 participants, corresponding to medium and large effect sizes [283, 284, 285, 286]. Specifically, while one study [284] recruited 21 participants, others recruited 5, 11, and 14 participants, respectively [283, 285, 286]. The number of subjects in this study, therefore, falls within the range observed in comparable research and aligns with the methodological rigor expected for studies with these experimental setups. This study is well-positioned to provide meaningful insights into user preferences using physiological and eye-tracking signals, even with a smaller sample size, due to the within-subject design that enhances statistical power and sensitivity. However, to improve the generalizability of this study further, future research should aim to recruit a larger and more diverse participant pool. Expanding the sample size will help capture a broader range of individual differences, ensuring that the findings can be applied more widely across different demographics and use cases.

7.1.4 Data Collection Procedure

7.1.4.1 Traffic Situation Recording System

All traffic situations were recorded under naturalistic driving conditions using the action camera mounted on the dashboard as shown in Figure 60, and video clips were judiciously trimmed to durations ranging from 30 to 75 seconds. The trimming process was tailored to meet the minimal time required for participants to comprehend each traffic situation effectively.

Figure 60. AKASO V50 camera mounted on the dashboard



7.1.4.2 Survey Procedure

Each participant received an invitation email through Prolific. To participate in the online survey, they accessed the survey link via the Prolific server and were asked to carefully read the survey information and provide their consent on a consent page. The online survey consisted of two sections: a one-time questionnaire and a situation questionnaire, as shown in Table 20. After providing their consent, participants completed the one-time questionnaire before beginning the first session, as illustrated in Figure 63. At the beginning of each session, each video clip of 18 traffic situations was shown before the questions, and the participants could watch the clip in the survey window. After completing all situation questionnaires, participants were asked to read a feedback letter for any missing information, how to withdraw participation, data retaining years, and a completion code for finalizing their survey completion in Prolific.

Figure 63. Video session procedure



Each response was saved in the Qualtrics server, and their response quality was evaluated by the first question of the situation questionnaire. The first Multi-Choice Question (MCQ) was to select the correct choices in each video clip for quality evaluation of responses. For scoring-based evaluation, three participants who scored below zero were excluded (Score: correct answer = +1/ wrong answer = -2). As shown in Table 20, we designed this online survey with multiple questions and selected a few responses as variables to indicate individual perceptions of self-driving for each situation. The preference questions regarding the binary preference for self-driving and its five preferred actions originated from Park et al. (2020) [286]. The binary preferences are 'Prefer' and 'Not prefer.' The five preferred actions are: 'Act-Inform,' 'Inform-Act,' 'Act-No Inform,' 'Inform-Consent,' and 'Alert-Handover.' Other responses in Table 20 were used as variables to understand individual preferences for self-driving and associations with preference/preferred actions.

Table 20. Details of one-time and situation questionnaires

One-time questionnaire	Responses
Age	23~44 / 60+ years old
Gender	Men / Women / Not Identified
Years of driving [years]	2~28 (Y-M) / 30+ years (Old)
Approx. total mileage [km]	5000~1140000
Advanced features experienced (MCQ)	Select all checked in 0~7 functions
Trust level on self-driving [7]	Scale 1-7
Potential trust level [5]	Scale 1-5
Situation questionnaire	Responses
Details of a situation [4~5] (MCQ)	Select all situation details that matched (Depending on each situation)
Preference for self-driving mode [2]	Prefer (1) and Not prefer (2)
Preferred actions [5]	5 options in Table 18
Reasons for preference choice (short answer)	Participant's text responses.
Difficulty level of driving [5]	Scale 1-5
Two influencing factors of your preference choice [5] (MCQ)	Five factors: Traffic flow, pedestrians, driving vehicle behaviours, etc.
- [] is the number of Likert scales. - MCQ: multiple choice question.	

7.1.4.3 Experiment Procedure

Physiological signals and eye-tracking data recording are conducted while participants watch the traffic situation videos during each session. These recorded data are then used to extract features. Additionally, NASA-TLX responses are collected from participants at three key points: before the start of the first session, after half of the 18 sessions (i.e., after the 9th session) and after the final session. These responses serve as a tool for the subjective measurement of workload that will be utilized in the chapter focused on workload prediction.

7.2 Data Analysis

In order to compare results between the Y-M and the Old group, we replicated the methodologies of data processing and statistical analysis as outlined in [241].

7.2.1 Survey Analysis

In order to compare results between the Y-M and the Old group, we replicated the methodologies of data processing and statistical analysis as outlined in [241]. The tools employed for this purpose included MATLAB, G power and IBM SPSS [287, 288, 289]. A comprehensive overview of the statistical tests utilized in this study based on variable types can be found in Table 21. Additionally, Table 22 provides a baseline power analysis presenting sample numbers, degrees of freedom, effect size and power (0.8). This rigorous approach to statistical analysis ensures the reliability and validity of our findings by maintaining consistency with established methodologies and tools. It also facilitates a meaningful comparison between the results of the current study and those obtained in the previous research, particularly concerning the Old group data.

Table 21. Statistical tests were used in this study.

Statistical test	Ob.	Variables	
Chi-square	170, 1020, 1190	Traffic complexity [2]	Difficulty level [5]
			Preference mode [2]
			Difficulty level [5]
		Preference mode [2]	Preferred actions [5]
Kruskal-Wallis, Correspondence analysis	1190	Preferred actions [5]	Difficulty level [5]
Correspondence analysis	1190	Situation types [14]	Preferred actions [5]
Spearman's rank correlation analysis	85	the numbers of 'prefer' and feature experienced, current trust level	

- Ob.: Number of observations

- [] is the number of variable levels.

Table 22. Baseline of statistical power analysis (Power = 0.8, Effective size $w = 0.25$ or 0.3).

Test	Situation	df	α err prob.	β err prob.
X^2	Toronto/Waterloo and roundabout (Sample size = 1190)	1	7.16e-15	0.2 (p -value requires much lower than the α err prob to ensure a statistical significance)
		4	5.53e-13	
	Situation-by-situation (Sample size = 170)	1	0.0021	
		4	0.016	
Correlation	Trust, numbers of prefer, and features experienced. (Sample size = 85)	N/ A	0.024	
	Toronto/Waterloo and roundabout (Sample size = 1190)	N/ A	1.00e-10	

- w : Cohen's w - X^2 : Critical X^2
- λ : Noncentrality parameter - α err prob.: the probability of making a type I error

Feature engineering for survey data

In Table 23, all features and their description are presented. These features were categorized into three types: (1) background features (demographic/trust), (2) situation response features, and (3) traffic condition features. The background features were collected once per participant, offering insights into personal characteristics. Situation responses were collected based on eighteen traffic situations for each participant. Most situation responses and their scales are straightforward, as explained in Table 23. However, there are a few responses, such as difficulty level, pedestrian level and traffic complexity that may not be straightforward. The details of these features are below.

- Difficulty level (integer scale 1-5): it is a subjective rating for each traffic situation. The usage of the ML feature is normalized based on min./max. rating of each participant.
- Traffic complexity (values 1 or 2): Depending on cities, e.g., in Waterloo (1) or Toronto (2).
- Pedestrian levels (integer scale 1-4): it is based on both proximity between pedestrians and the driving vehicle and potential risks on pedestrians, e.g., 'highway to city' has no pedestrian (1), 'left turn' has a minimum risk (2), 'Multi-situation' has a low pedestrian risk (3) and 'Parking lot' has a high risk due to pedestrians walking freely (4).

The influencing factors are based on the number of responses to each factor. Categorical variables were converted into features for ML algorithms using the one-hot encoding technique except for interaction terms. A simple multiplication technique was used to create the interaction terms. The responses to influencing factors such as traffic flow, pedestrians, behaviours, etc., were multiplied by the normalized difficulty level.

$$\text{Interaction term}_1 = \text{Influencing factor}_1 \times \text{difficulty level}$$

It captured both non-linear and synergistic effects between the pairs, capable of either enhancing or diminishing the effects of influencing terms. For instance, if a participant responds to a low value of an influencing factor compared to other participants and is rated with a difficulty level below 3, presumably the max level is 5, it can diminish the value of interaction terms. All these features can be integrated into ML models to identify patterns and enable data-driven decisions customized for each participant.

7.2.2 Experiment Data Analysis

As we introduced earlier in Chapter 2.4 regarding user preferences, these preferences can be influenced by two main time frames. First, user preferences can be shaped by initial judgments such as first impressions. Secondly, they can be re-evaluated and adjusted by subsequent stimuli, environmental changes or varying contexts as post-preferences. Given the dynamic nature of traffic situations, user preferences may fluctuate due to various factors in each scenario. Therefore, our epoch extractions focus on two domains: the first impression and the main event in each traffic situation. Additionally, it is crucial to provide users with sufficient time for decision-making regarding their preferences immediately following the first impression and the main event. This ensures that physiological signals and eye-tracking data can be extracted over a sufficiently long epoch duration. For both the first impression and the main traffic event, we extract two epochs in each scenario as illustrated in Figure 64.

Figure 64. Two epoch durations for first impression and main traffic events (Frames are extracted from ‘S1: Police in high traffic volume, Toronto’)

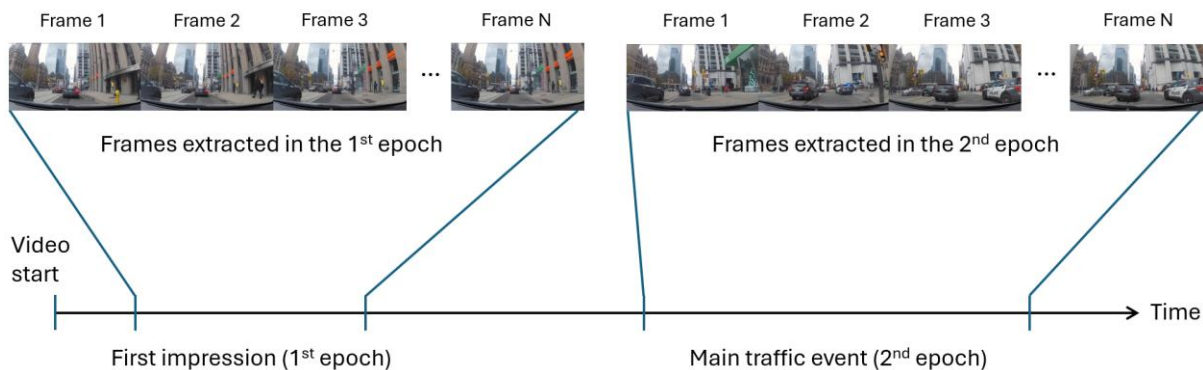


Table 23. Features extracted from the survey responses and the traffic situation information (All scale is integer before applying it into ML models as features).

Features		Scales	Feature names
Background features	Gender	1: Men / 2: Women	'gender'
	Age range	(Younger) 23~44 years old / (Older) 60+ years old	'age'
	Driving experience	1~40+ years (closely related to age)	'years'
	Number of advanced functions experienced	0-7: how many functions were experienced before	'n_features'
	Trust level on self-driving (intrinsic response)	1-7 scale: how much you trust self-driving	'trust'
Situation features	Difficulty level (subjective rating)	1-5: how difficult to understand/perceive potential risks	'difficulty'
	Traffic speed	1-3 scale: Low / Medium/ Fast	'speed'
	Influencing factors	Number of the responses on traffic flow	'facts_traffic_flow'
		Number of the responses to pedestrian	'facts_ped'
		Number of the responses on vehicle behaviours	'facts_behaviour'
		Number of the responses on traffic unexpectedness	'facts_unexpect'
	Interaction terms: Combinations of two features	Difficulty level and the factor, 'facts_traffic flow'	'interac_traffic_flow'
		Difficulty level and the factor, 'facts_ped'	'interac_ped'
		Difficulty level and the factor, 'facts_behaviours'	'interac_behaviour'
		Difficulty level and the factor, 'facts_unexpect'	'interac_unexpect'
Traffic conditions	Average preference rate	Number of 'Prefer' choices in each situation	'pref_rate_sc'
	Pedestrian level in each situation	1-4 (1: no pedestrian/ 2: no risk / 3: minimum/ 4: high)	'pedestrian_level'
	Traffic complexity	1-2 (low ~ Waterloo and high ~ Toronto)	'city'
	Vehicle counts per hour in each area	Averaged number of vehicles. [vehicles/hour]	'traffic_counts'
	Number of lanes	1-3: How many lanes were while driving	'n_lanes'
	Divider	0 or 1: Road divider	'divider'
	Number of obstacle types	0-2: e.g., construction, parked vehicles, etc.	'obstacles'

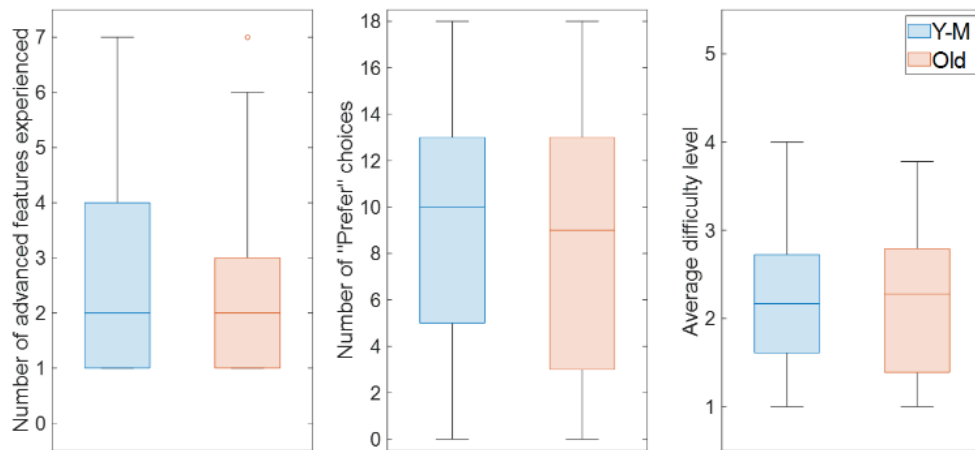
7.3 Survey Result

Both age groups exhibited high proportions, with over 85% of individuals using at least one of the advanced features. Due to the greater mileage and years of driving in the Old group, all participants in that group experienced at least one feature. However, the average number of features experienced was slightly higher in the young-to-middle-aged (Y-M) group. Table 24 provides the average and standard deviation of demographic and situation questions highlighting differences between the Y-M and Old age groups. Regarding trust levels, the Y-M group consistently demonstrated higher levels, with the most significant difference observed in past trust levels. This disparity may be attributed to the less favourable perception of self-driving among Old participants compared to the Y-M group. Moreover, the Old group extensively assessed their current trust in self-driving in comparison to the Y-M group, suggesting a more pronounced impact of self-driving and advanced vehicle technology on their inherent perception. In terms of the average number of preferences for self-driving, the Y-M group expressed a greater preference for self-driving in both traffic complexities, accompanied by a higher standard deviation in the Old group. The variation in preferences between the two traffic complexities may suggest that the Old group was less inclined to change their preferences based on traffic complexities and region types, such as Toronto (Metropolitan area) and Waterloo (city). Interestingly, both age groups recognized and similarly rated difficulty levels, indicating an apparent consistency in their assessments across the two traffic complexities. This comprehensive analysis provides valuable insights into the distinct traffic dynamics of Toronto and Waterloo, contributing to a deeper understanding of the situations. Figure 65 presents boxplots illustrating features experienced, the number of preferences and difficulty levels in the two age groups. In the Y-M group, the upper 50-75 percentile of the number of features experienced exhibited a more widely distributed pattern compared to the Old group. This distribution suggests that the Y-M either had more opportunities to use vehicles equipped with advanced features or were more willing to utilize such features, or possibly a combination of both factors. Regarding the number of preferences, both age groups displayed similar levels in the upper percentiles. However, the lower percentile of the Old group was more distributed toward the bottom. This distribution may indicate that some individuals in this percentile were strongly inclined toward manual maneuvering, irrespective of the rated difficulty levels in each traffic complexity. The insights from these boxplots contribute to a comprehensive understanding of the varying experiences and preferences within each age group.

Table 24. Average data for survey responses

List		Y-M	Old
Average number of experienced features		2.49 (1.51)	2.39 (1.63)
Average trust level on self-driving	Current	3.86 (1.33)	3.56 (1.69)
	Past	2.79 (1.46)	2.11 (1.33)
	Future expectation	3.99 (0.84)	3.72 (1.02)
Average number of preferences	High traffic comp.	3.74 (2.58)	3.53 (2.96)
	Low traffic comp.	5.42 (2.51)	4.82 (3.13)
Average difficulty levels	High traffic comp.	2.42 (0.81)	2.33 (0.92)
	Low traffic comp.	1.98 (0.71)	1.99 (0.73)
- Values shown are Mean (SD)		- N = number of participants	- comp.: complexity

Figure 65. Age differences in numbers of advanced features experienced (demographic information), number of ‘Prefer’ choices and average difficulty levels in 18 situations



7.3.1 Traffic Complexity Effect on Difficulty Levels

Our prior study demonstrated an association between traffic complexities and difficulty, indicating that participants rated higher difficulty levels in high traffic complexity and vice versa [188]. In this study, the Old group exhibited similar ratings on difficulty levels as the Y-M group on two traffic complexities in Table 24, appropriately gauging their difficulty levels. However, the Chi-square test assessing the relationship between these variables yielded no statistical significance when comparing p-values and alpha values based on the number of samples in the Old group. This is in contrast to the Y-M group that showed statistical significance in the previous study [188]. In this study, the Old group did not show a statistical significance between traffic complexity and difficulty levels (Toronto/Waterloo and

roundabout: $p = 6.957E-10$ over a required p -value $5.53e-13$, Transition situations, ‘Highway to city’ and ‘City to highway’: $p = .994$ and $.294$ over a required p -value $.0016$). Compared to the Y-M group result, the Old group exhibited a situation: the ‘Police’ situation in Table 25. Except for the ‘Police’ and ‘City to Highway’ situations, the Old group showed higher lambda values than the Y-M group in all other situations. This indicates that the perceived difficulty levels rated by the Y-M group have a greater impact on estimating traffic complexities than the levels rated by the Old group. Not only does the Y-M group rely more on the perceived difficulty levels when deciding on traffic complexity, but the Old group is less likely to associate difficulty levels with binary traffic complexities. This result indicates that perceived difficulty levels at various traffic situations can be a factor in designing a user perception-based system in the Y-M group.

Table 25. Results of the Chi-square test: Traffic complexity and difficulty levels

Situations (N = 170)	p		X^2		λ	
	Y-M	Old	Y-M	Old	Y-M	Old
Police	.024	.014*	11.251	12.455	.176 (.017)	.224 (.005)
Multi-situation	.060	.064	9.033	8.898	.224 (.018)	.200 (.091)
Turn Right	.003	.003	16.356	15.697	.294 (.009)	.235 (.045)
Turn Left	.001*	.267	18.301	5.199	.318 (.007)	.165 (.176)
Parking lot	.011*	.416	13.158	3.928	.212 (.092)	.141 (.228)
aroundCar	.000*	.008	22.872	13.707	.318 (.010)	.247 (.044)
Roundabout	.001*	.028	18.158	10.873	.294 (.021)	.224 (.074)
H2C	.422	.997	3.882	.146	.071 (.473)	.024 (.803)
C2H	.763	.294	1.849	4.932	.071 (.517)	.082 (.574)

* $p < .0016$

- Degree of freedom: 4.

- λ (Lambda): Traffic complexity dependent. (Approximate significance)

7.3.2 Traffic Complexity Effect on Preference Modes

Figure 66 illustrates the differences in preferences between low and high traffic complexity (volume) situations in the two age groups. Except for the 'Highway to city' situation in the Old group, both age groups exhibited positive differences in all situations, indicating a preference for more self-driving in low-traffic situations. Regarding age differences in such preference variances, the young-to-middle-aged (Y-M) group tended to discriminate their preferences for self-driving between two traffic complexities

except for the 'Multi-situation' situation. Additionally, the Old group displayed higher discrepancies in preferences between the two complexities for certain situations such as 'Parking lot,' 'aroundCar,' and 'Turn Right. To further understand the associations between traffic complexities and preferences, Table 26 presents the Chi-square test results for each situation. Analogous to the tests between complexities and difficulty levels, the Y-M group exhibited statistical significance and better error reduction probabilities in the majority of situations compared to the Old group. This indicates a greater propensity for the Y-M group to formulate their preferences in response to traffic situations than the Old group. Such findings augment our understanding of how preferences for self-driving vehicles diverge across age groups in response to varying traffic complexities and situations.

Figure 66. Proportion differences between the two traffic complexities (% difference = % Low traffic complexity – % High traffic complexity).

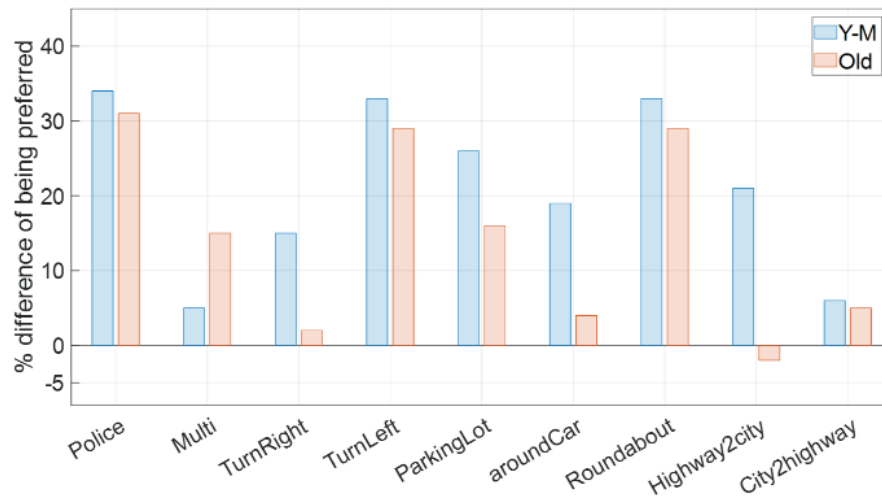


Table 26. Chi-square test: Traffic complexity and preference mode.

Situations (N = 170)	p		X^2		λ	
	Y-M	Old	Y-M	Old	Y-M	Old
Police	.000*	.000*	20.7	17.691	.164 (.231)	N/A (c)
Multi-situation	.538	.046	0.38	3.985	N/A (c)	.111 (.328)
Turn Right	.000*	.000*	18.5	14.780	.305 (.006)	.241 (.037)
Turn Left	.001*	.031	11.47	4.664	.192 (.101)	.066 (.587)
Parking lot	.018	.713	5.623	.136	N/A (c)	N/A (c)
aroundCar	.014	.645	6.044	.212	.138 (.231)	.035 (.745)
Roundabout	.000*	.000*	19.90	15.005	.081 (.587)	.178 (.158)

H2C	.754	.758	0.098	.095	N/A (c)	N/A (c)
C2H	.343	.506	0.899	.443	N/A (c)	N/A (c)

* $p < .0021$

- Degree of freedom: 1

- λ , Lambda: Preference mode dependent. (Approximate significance)

- c: cannot be computed because the asymptotic standard error equals 0.

7.3.3 Statistical Tests: Binary Preferences and Difficulty Levels

As shown in Table 27, in Toronto/Waterloo situations, the relationship between binary preferences for self-driving and difficulty levels was statistically significant in both age groups, Y-M and Old, but not statistically significant in roundabout situations. Between these two age groups, the average and standard deviation of difficulty level exhibited no significant difference in the two preferences, 'Prefer' and 'Not prefer.' Both age groups rated difficulty levels between 1 and 3 over 90% in situations where they preferred self-driving. The proportions of difficulty levels decrease from level 1 to 5 as the level increases in the 'prefer' choices. However, this trend is not evident in the 'not preferred' choices. Instead, a similar trend is observed from either the proportions of level 2 or 3 in the 'not prefer' choices. The average difficulty level for both 'prefer' and 'not prefer' choices is around 1.9 and 2.5, respectively. From a macroscopic perspective, the bifurcation of difficulty levels for the two preference modes is either at level 2 or 3. This implies that these difficulty levels could be the most challenging to predict when determining driver preferences.

Table 27. Chi-square test results: mode preference and difficulty levels.

Traffic situations (14 situations)			Difficulty levels						p
			1	2	3	4	5	Avg (SD)	
Y-M	Cities (N = 1020)	Prefer	44.5	31.5	15.2	6.9	1.8	1.90 (1.06)	.000*
		Not prefer	22.4	21.8	30.2	17.4	8.2	2.67 (1.23)	
	Roundabout (N = 170)	Prefer	47.2	31.5	15.7	4.6	0.9	1.81 (0.93)	.004
		Not prefer	19.4	38.7	27.4	11.3	3.2	2.40 (1.03)	
Old	Cities (N = 1020)	Prefer	44.4	29.3	19.9	6.0	0.5	1.89 (0.96)	.000*
		Not prefer	29.2	20.4	32.0	14.3	4.1	2.44 (1.17)	
	Roundabout (N = 170)	Prefer	49.5	29.9	13.4	7.2	0.0	2.13 (0.94)	.002
		Not prefer	30.1	21.9	32.9	12.3	2.7	2.36 (1.12)	

* $p < 5.62e-11$ (N = 1020, Cities) / * $p < .0016$ (N = 170, Roundabout)

- Value shown in each cell is a proportion value (%).

7.3.4 Preference Modes and Preferred Actions

While the Y-M group demonstrated statistical significance between preference modes and preferred actions through the Chi-square test (Cities: $p = 6.35e-89$ below a required p -value $5.53e-13$ and Roundabout: $p = 1.51e-14$ below a required p -value $.0016$), the Old group similarly showed statistical significance (Cities: $p = 6.35e-89$ below a required p -value $5.53e-13$ and Roundabout: $p = 1.51e-14$ below a required p -value $.0016$) in Table 28. Figure 67 illustrates overall preferences for the 'being informed' action and human-engaged actions. The Old group slightly less preferred the AV action of being informed than the Y-M group, but both age groups preferred over 80% of the action regardless of traffic complexities. This preference might stem from the design of the preferred actions that included 'being informed' action amidst the total of five actions (80% of actions included 'being informed'). In a situation-by-situation analysis, certain situations at low traffic complexity, such as 'Police,' 'Turn Right,' and 'City to highway' situations, showed higher discrepancies between the two age groups than other situations in Figure 68. This indicates age differences in the 'being informed' action between these age groups. It was found that the action 'being informed' was highly preferable to the Y-M group, resulting in positive values of the preference difference in the action (except for the high traffic complexity of the 'City to highway' situation). This preference could be explained by the fact that Y-M users either want to interact with the AV decision-making system or prefer to engage in vehicle control after checking risky or high-traffic complexity environments. Additionally, the preference differences of the action between the age groups decreased in high-traffic complexity situations, suggesting that both age groups showed fewer discrepancies in their preference for the 'being informed' action in such situations. Between human-engaged and -disengaged action groups, the overall differences between such action groups in the two complexities were lower in the Old group, indicating that the Old group had less impact on choosing preferred actions based on traffic complexity, as depicted in Figure 69. In situation-by-situation analysis, the Y-M group preferred human-engaged actions more than the Old group in high traffic complexity.

Table 28. Chi-square test results: preferred actions of high/low traffic conditions in each group. (%)

Scenario types		Preferred actions					P-value
		Act- Inform	Inform- Act	Act-No Inform	Inform- Consent	Alert- Handover	
Cities (N = 1020)	'Prefer'	37.8	24.4	18.0	17.1	2.8	.000*
	'Not prefer'	5.5	10.8	1.2	56.3	26.3	6.35e-89
Roundabout	'Prefer'	38.0	22.2	21.3	15.7	2.8	.000*

(N = 170)	'Not prefer'	3.2	8.1	4.8	62.9	21.0	1.51e-14
14 scenarios		Group 1 (Human-disengaged)			Group 2 (Human-engaged)		P-value
14 scenarios	'Prefer'		80.4		19.6		.000*
(N = 1190)	'Not prefer'		17.3		82.7		7.3e-103

- 14 scenarios included both 'Cities' and 'Roundabout' scenarios.

- Value shown in each cell is a proportion value (%).

Figure 67. Preferences for 'being informed' action and human-engaged actions in the two age groups (Human-engaged actions included 'Inform-Consent' and 'Alert-Handover').

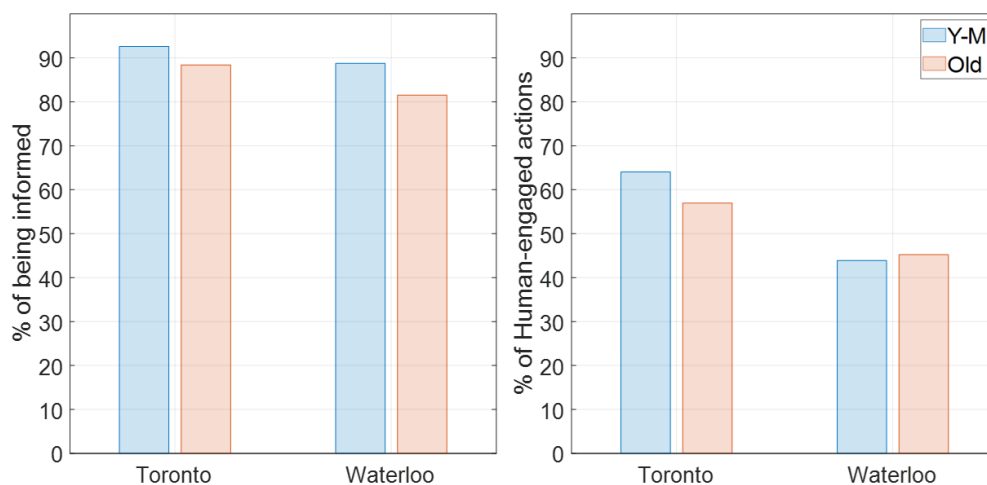


Figure 68. Proportion differences of being informed between the age groups in two traffic complexities (difference % = Y-M group – Old group).

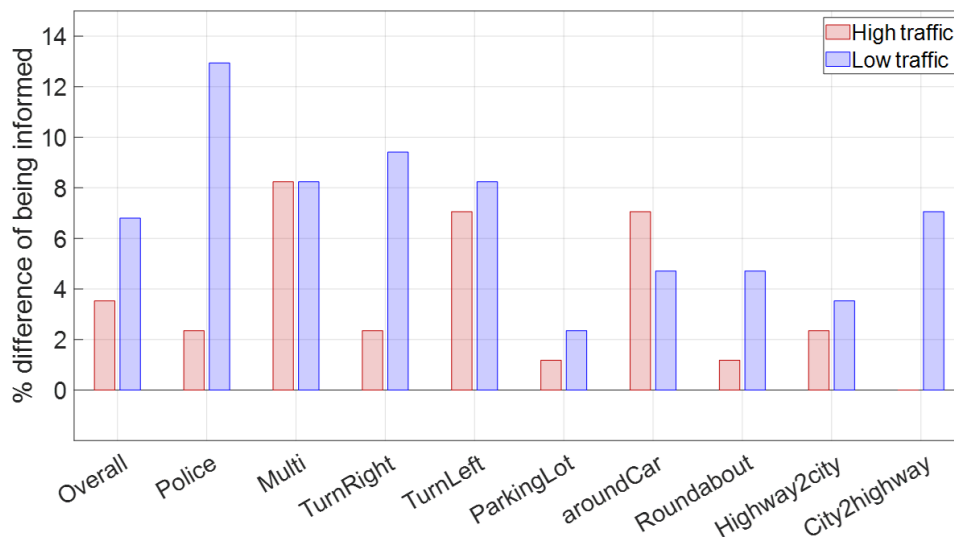
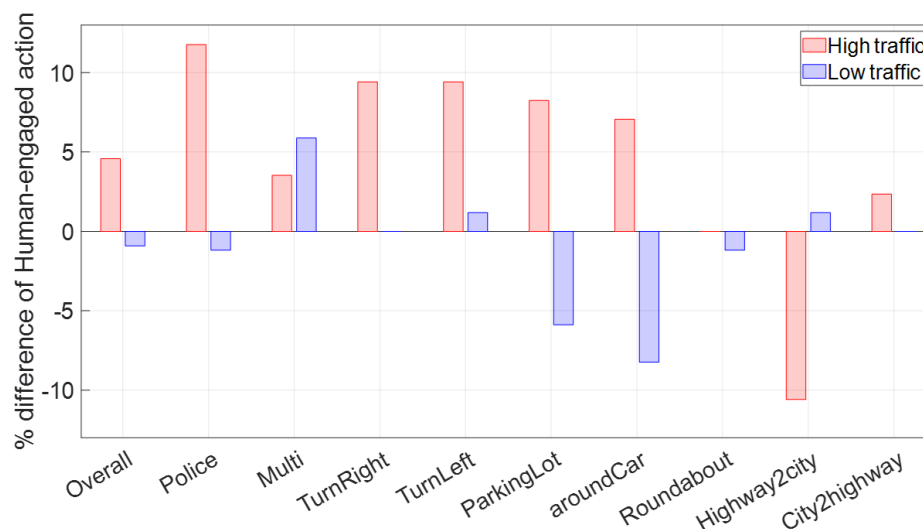


Figure 69. Proportion differences of human-engaged actions between the age groups (% Y-M – Old) in two traffic complexities.



The Old group preferred human-engaged action slightly more than the Y-M group in certain situations with low traffic complexity, such as 'Parking lot' and 'aroundCar.' The group only preferred human-engaged action more than the Y-M group in the 'Highway to city' situation in high traffic complexity. In an attempt to better understand the effects of age difference on preferred actions, correspondence analysis was performed. Figure 70 illustrates the change in the two age groups' preferred actions in each situation. The arrowheads represented the Old group, and the colours of the arrow lines, red and blue, represented high and low traffic complexity, respectively. In certain situations, including the traffic complexity effect, the age variable could significantly impact changing preferred actions either between actions in the same action group, human-engaged and -disengaged or between such action groups. The situations moving to another action group by the age effect were shown in a total of four situations (three high-traffic and one low-traffic complexity situation). The other three situations remained in the central areas between the two action groups, except for the roundabout at the low traffic volume. The changes of preferred actions on each side were highly dependent on situation types approaching either closer to a certain type of action or to neutral regions between these actions. The formal cases were shown at both traffic complexities of 'Parking lot' situations, 'aroundCar' and 'Turn Right' at low traffic complexity. Specifically, both 'Parking lot' situations were highly moved to 'Alert-handover' as explained by short answers from participants foreseeing a high chance of unexpected risks and situations in parking lots regardless of traffic complexity. In addition, in the low-traffic 'Turn Right' situation, the Old group preferred the 'No Inform' action that was highly correlated with the human-disengaged action. Otherwise, the Y-M group preferred 'being informed' in the situation.

Current trust level	.47**	.23	1		
Past trust level	.36**	.21	.56**	1	
Potential trust level	.49**	.27*	.53**	.34**	1
Variables	Old: 60+ years old				
	V1	V2	V3	V4	V5
Num. of 'Prefer'	1				
Num. of experienced functions	-0.08	1			
Current trust level	.43**	.127	1		
Past trust level	.34**	.30*	.55**	1	
Potential trust level	.27**	.009	.68**	.41**	1

*: $p < .024$ / **: $p < .0024$

- Num. of 'Prefer': Number of 'Prefer' choices
- Num. of exp.: Number of experienced features
- Past trust lv: past trust levels when participants first heard of AVs
- Potential trust lv: potential of trust level change in future

7.3.6 Interaction Effects: Preference and Individual Perception/Experience

In this section, the only independent variable available for analysis in each traffic scenario is the difficulty level of driving in that specific scenario. This independent variable alone does not allow for the study of interaction effects with other variables. However, by using the "Num. of 'Prefer'" (Number of preferred scenarios for self-driving) as the dependent variable, interaction effects among the other variables included in Table 30 can be analyzed. Below, the results of the comparative analysis of interaction effects between the Y-M group and the Older group are presented. The younger group (Y-M) demonstrated significant interaction effects across multiple trust and experience-related variables, indicating a complex decision-making process. The interaction between Current Trust and Future Expectations (IE 2) was highly significant ($p < 0.001$), emphasizing the importance of how trust influences future-oriented decisions. Additionally, interactions involving past trust, such as Past Trust $\times$ N. Exp. (IE 3) ($p = 0.016$), suggest that younger adults heavily rely on past experiences when forming preferences. More complex interactions, such as Past Trust $\times$ Current Trust $\times$ Future Expectations (IE 6) ($p < 0.001$) and Past Trust $\times$ Current Trust $\times$ N. Exp. (IE 7) ($p = 0.008$), further highlight the dynamic interplay of trust and experiential factors for this group. In contrast, the older group showed fewer significant interaction effects, reflecting a simpler decision-making framework. The interaction between Current Trust and Future Expectations (IE 2) ($p = 0.011$) was one of the few significant results, indicating that future expectations remain a key factor for this group. However, interactions involving experience-related variables, such as N. Exp., were generally not significant. For example, Past Trust $\times$ N. Exp. (IE

3) and Future Exp. $\times$ N. Exp. (IE 5), which were significant for the younger group, were not significant for the older group, suggesting that older adults may place less emphasis on experiential factors. The findings reveal distinct decision-making patterns across the two age groups. Younger adults rely on a more dynamic and multifaceted approach, integrating trust and experience-related factors in both simple and complex ways. In particular, their sensitivity to higher-order interactions, such as those involving multiple trust variables and experiential factors (e.g., IE 6 and IE 7), suggests a strong adaptability in their decision-making process. On the other hand, older adults focus on fewer, more straightforward interactions, with an emphasis on future expectations and current trust. This difference aligns with the idea that older adults may prefer clarity and stability in decision-making over complexity. Certain interaction effects, such as Past Trust $\times$ Future Exp. $\times$ N. Exp. (IE 9), were not computable for either group, likely due to insufficient data or unbalanced distributions across variable combinations. This limitation underscores the need for future studies to ensure balanced datasets and adequate sample sizes to capture higher-order interactions effectively. Addressing these limitations could provide a more comprehensive understanding of the interplay between trust and experiential factors.

Table 30. Significance results of interaction effects between the independent variables and the dependent variable shown in Table 21 (N = 85 for each variable)

Interaction Effect (IE #)	Independent Variable (IV #) : Dependent variable: # of preferred scenarios				Significance	
	IV 1	IV 2	IV 3	IV 4	Y-M	Older
IE 1	Past Trust	Current Trust	-	-	.271	.141
IE 2	Current Trust	Future Exp.	-	-	.000*	.011*
IE 3	Past Trust	N. Exp.	-	-	.016*	.429
IE 4	Current Trust	N. Exp.	-	-	.000*	.136
IE 5	Future Exp.	N. Exp.	-	-	.002*	.383
IE 6	Past Trust	Current Trust	Future Exp.	-	.000*	.134
IE 7	Past Trust	Current Trust	N. Exp	-	.008*	.856
IE 8	Current Trust	Future Exp.	N. Exp	-	.006*	.128
IE 9	Past Trust	Current Trust	Future Exp.	N. Exp.	N/A	N/A

- N. Exp.: Number of Experienced Features - Future Exp.: Future Trust Expectation

- Score Exp.: Score of experienced features

- N/A: no interaction effect due to one or more issues among the list below.

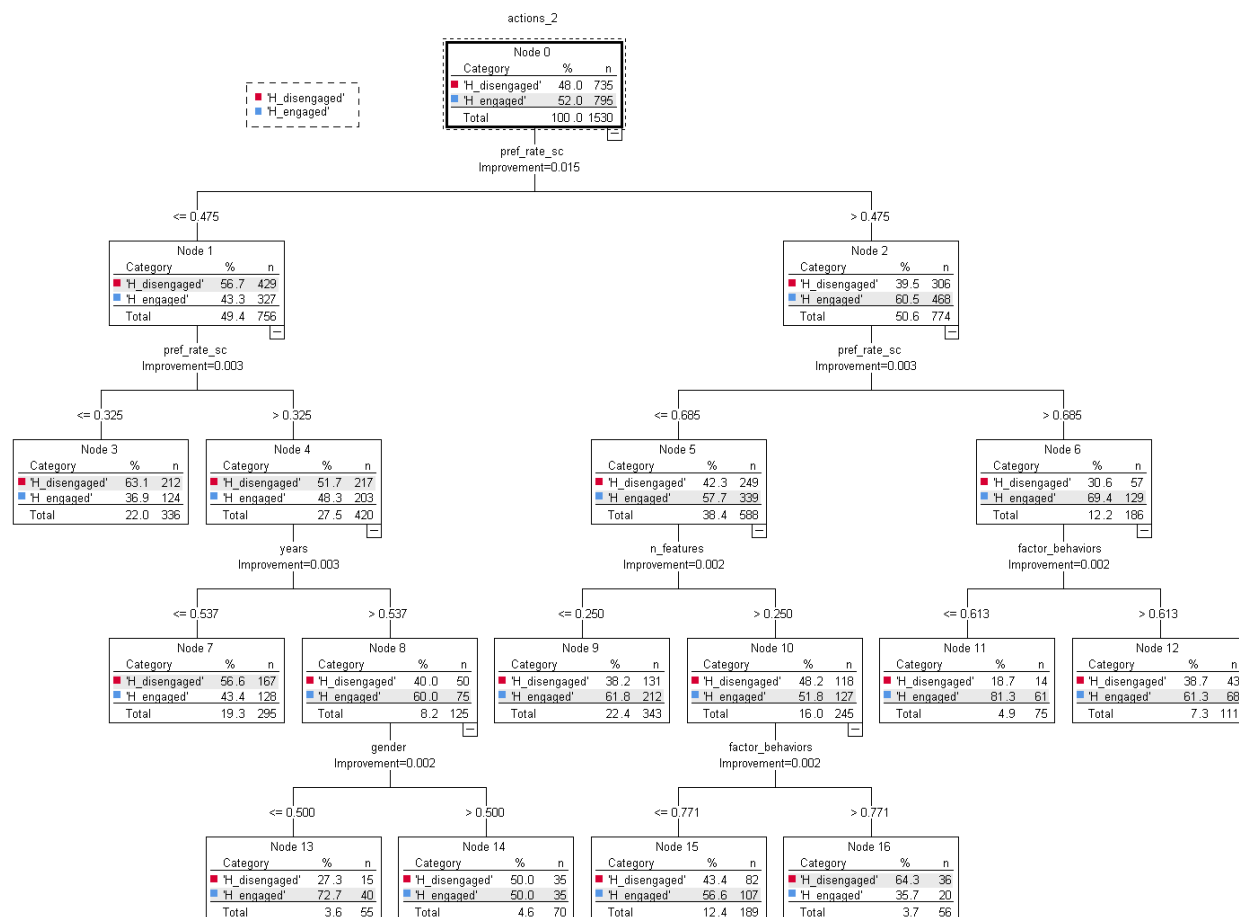
There are varied reasons why some of the interaction effects between variables are invisible in the table as follows: (1) Degree of Freedom, (2) Multicollinearity or overlap between variables, and (3) Model complexity. Under these issues, the invisible interaction effects are not estimable or computable due to insufficient data variability, a lack of unique contributions from the variables, or excessive model complexity that exceeds the dataset's capacity to support higher-order interactions. These limitations highlight the importance of ensuring balanced data distribution, reducing multicollinearity, and simplifying models to focus on interpretable and statistically meaningful effects.

7.3.7 Interaction Effects: Features Extracted from Survey Responses

In Section 7.4, regarding the survey-based preference model, I include feature importance analysis results derived from an algorithm built on game theory. The importance analysis provides an opportunity to optimize features for the model, which can enhance model efficiency and overall performance. In addition to the game theory-based feature importance analysis, this section briefly explores key features using statistical methods. To develop the user preference model based on survey responses, extracted features are described in Table 23. Among these features, to investigate the interaction effects of key features, I utilized a decision tree with the CRT (Classification and Regression Tree) growing method. The features assigned to the nodes were analyzed, and the results were compared between the Y-M group and the Older group. For the binary preference, the decision tree analyses for the Y-M group (ages 23–44) and the Older group (ages 60+) reveal distinct patterns in the factors influencing engagement, H_disengaged vs. H_engaged (Human-disengaged: usage of self-driving features and Human-engaged: manual control-based options) shown in Figure 71 and Figure 72, respectively. For the Y-M group, the primary splitting variable was ‘pref_rate_sc’ (Preference rate), emphasizing its central role in distinguishing engagement levels. Secondary variables, such as ‘years’ (Years of driving), ‘n_features’ (Number of advanced driver assistance features experienced), and ‘facts_behaviours’ (Number of the responses on vehicle behaviours), further segmented the data, suggesting broader interactions between these factors. Gender differences appeared at deeper nodes, but their overall contribution was minimal, as indicated by low improvement values. These results suggest that younger individuals prioritize preference-related factors when determining engagement. In contrast, the Older group showed a different hierarchy of variables, with years being the most influential splitting variable. This highlights the importance of temporal or experiential factors for older individuals. Secondary variables, such as ‘pref_rate_sc’, ‘facts_pedestrian’, and ‘Trust’, played significant roles, with larger improvement values observed at higher-level nodes. Although gender differences also appeared in the Older group, their overall impact remained low, similar to the Y-M group. The deeper splits and targeted patterns in the

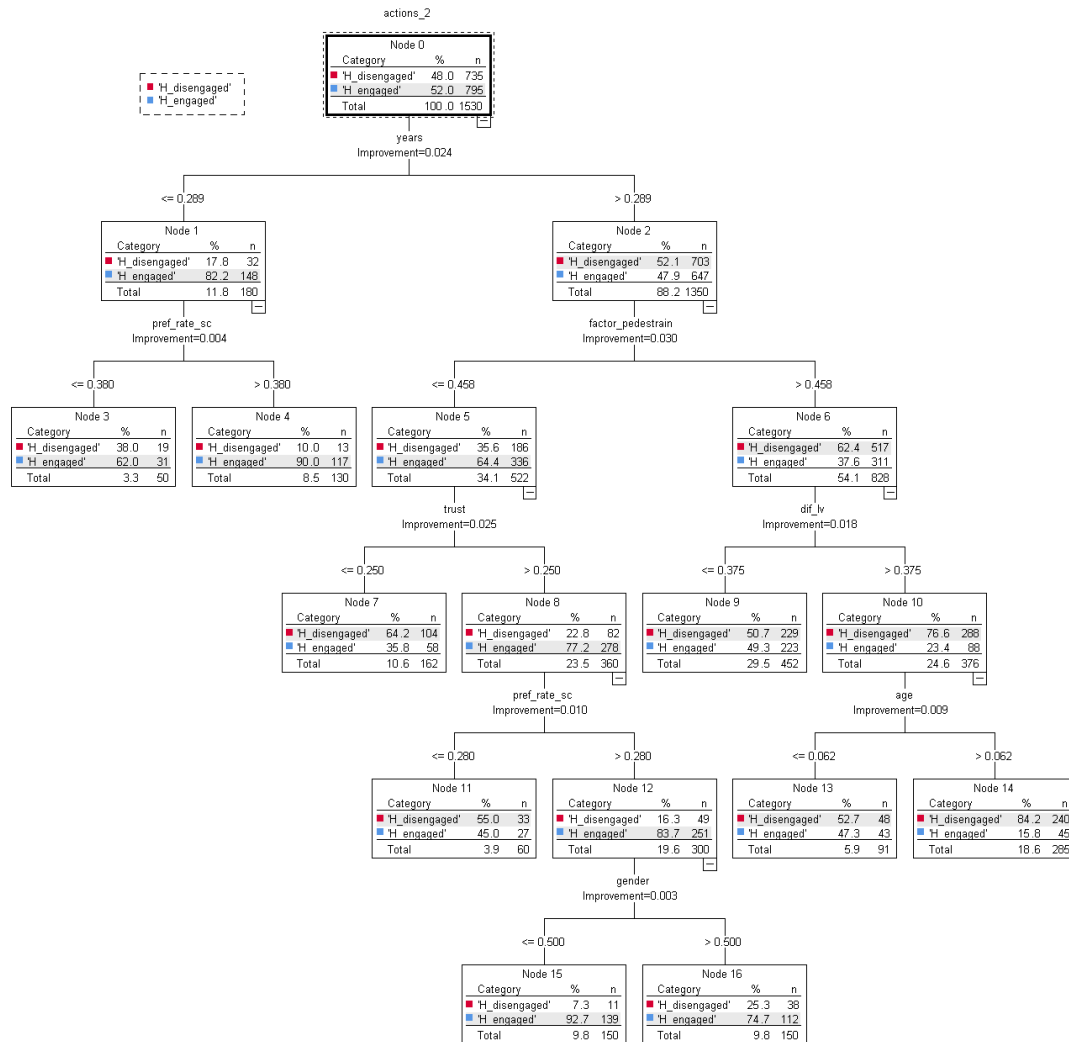
Older group suggest that engagement in this demographic is shaped by more specific conditions, such as trust and difficulty level.

Figure 71. Decision tree analysis for the Y-M group (Dependent variable: Binary preference)



When comparing the two groups, the primary and secondary variables differ significantly. While 'pref_rate_sc' dominates in the Y-M group, reflecting a focus on preference-based factors, 'years' takes precedence in the Older group, emphasizing the role of experience. Additionally, the Y-M group demonstrated more diversified secondary variables, whereas the Older group exhibited more targeted and condition-specific patterns. These differences highlight the need for tailored approaches to engagement strategies based on demographic characteristics. Future research should investigate these patterns further, integrating interaction effects and longitudinal analyses to refine predictive models for different age groups.

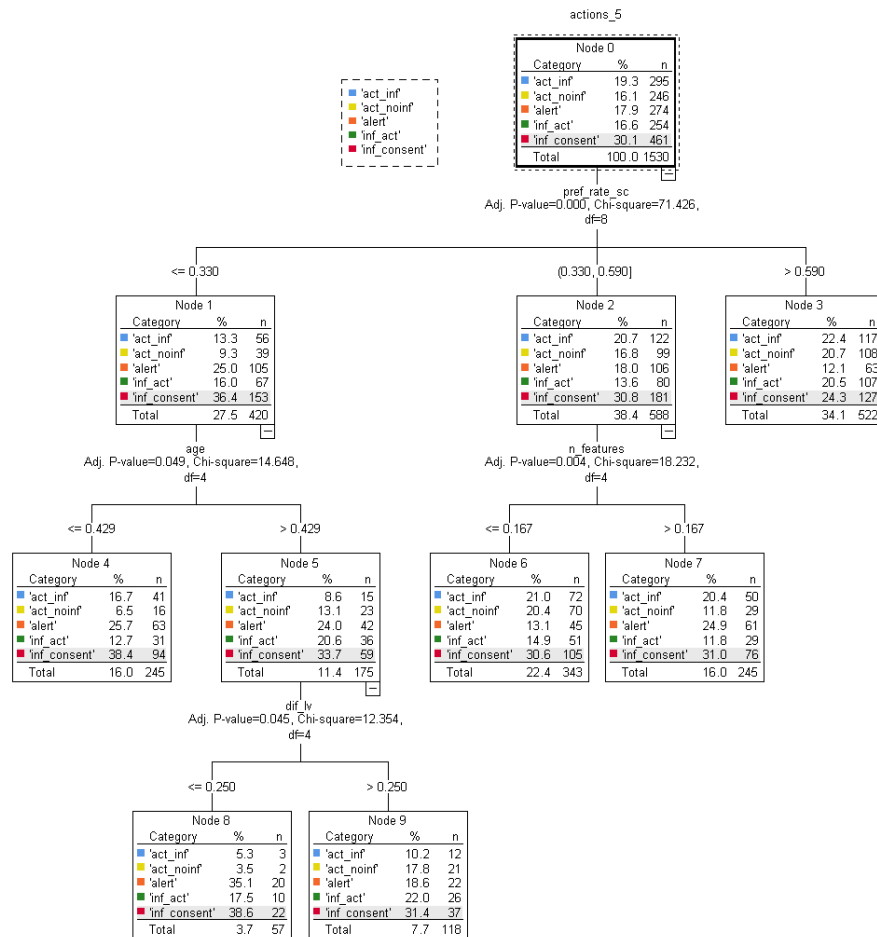
Figure 72. Decision tree analysis for the Older group (Dependent variable: Binary preference)



In the analysis of user preference for five self-driving actions, the decision tree with the growing method, CRT, does not show any node development in the Older group's feature dataset. Instead, the decision tree with the growing method, CHAID (Chi-squared Automatic Interaction Detection), was able to show the decision tree results in both the Y-M and Older groups. The decision tree analyses for the Y-M group (ages 23–44, Figure xx) shown in Figure 73 and the Older group shown in Figure 74 (ages 60+, Figure xx) illustrate distinct patterns in the influence of independent variables on predicting five self-driving actions (act_inf: Act-Inform, act_noinf: Act-No Inform, alert: Alert-Handover, inf_act: Inform-Act, and inf_consent: Inform-Consent). These results provide valuable insights into the decision-making processes and preferences of the two age groups. For the Y-M group, 'pref_rate_sc' emerged as the most influential variable, serving as the primary split at Node 0, with a significant adjusted p-value of 0.000 and a chi-square value of 71.426. This finding highlights the critical role of preference ratio in shaping

decisions among younger participants. Secondary variables, such as ‘age’, ‘n_features’, and ‘dif_lv’ (perception of difficulty level for driving in a scenario), were identified as key contributors in subsequent splits, with adjusted p -values ranging from 0.004 to 0.049. These results indicate that younger individuals tend to evaluate self-driving actions based on a combination of preference ratio, experience in advanced driver assistance features and age factors. The tree for the Y-M group exhibited a relatively balanced and straightforward structure, reflecting a preference-driven and practical decision-making approach.

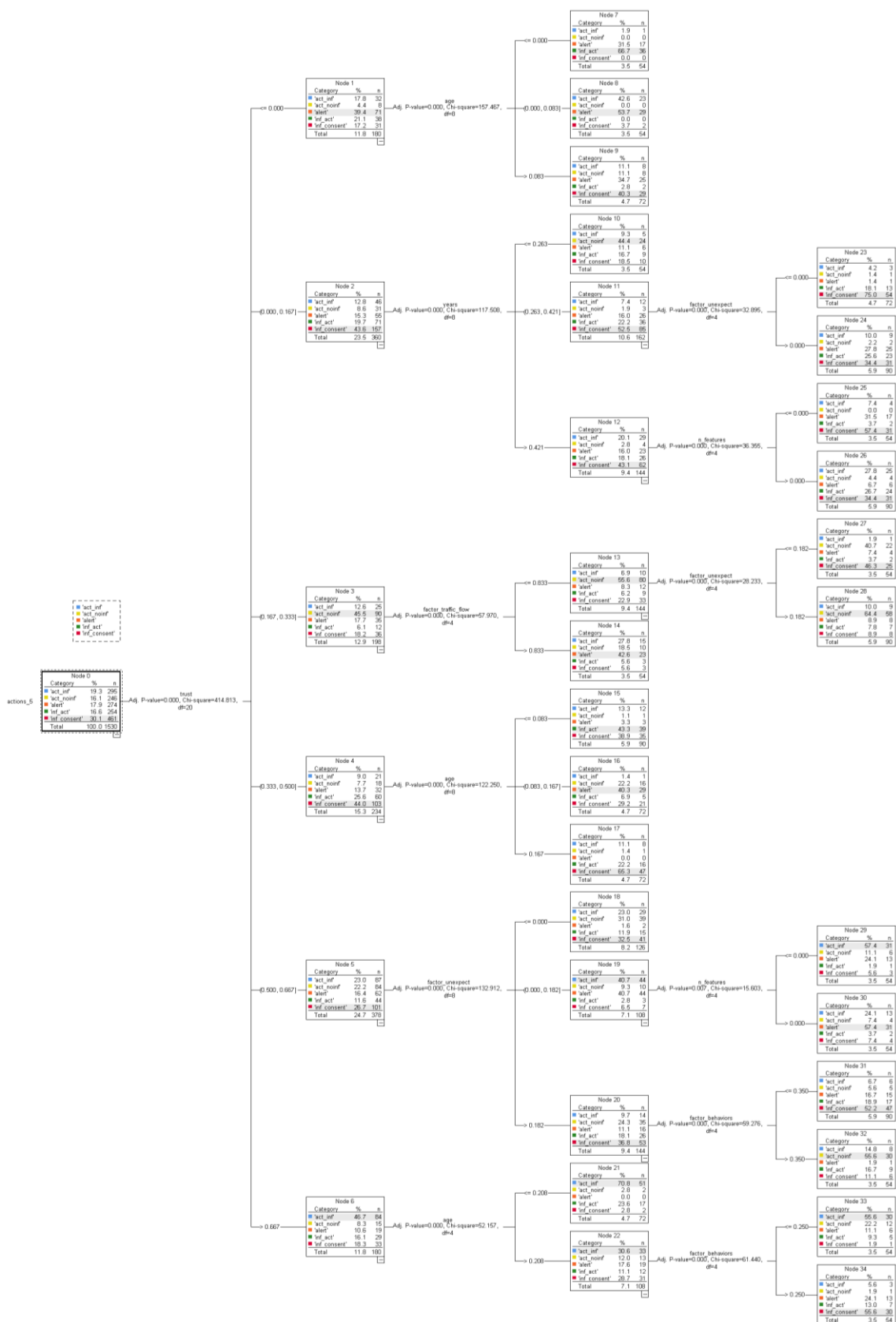
Figure 73. Decision tree analysis for the Y-M group (Dependent variable: Multiple actions)



In contrast, the Older group's decision tree shown in Figure 74 prioritized trust factor as the primary splitting variable at Node 0, with a significant adjusted p -value of 0.000 and a chi-square value of 157.487. This suggests that participants in the Older group place greater importance on the trust factor when evaluating self-driving actions. Secondary variables, such as ‘facts_traffic_flow’, ‘facts_behaviours’ (Number of the responses on vehicle behaviours), ‘facts_expected’ (Number of the

responses on traffic unexpectedness), and ‘years’, were critical in refining the classifications at deeper nodes, with adjusted p-values as low as 0.000 and chi-square values indicating strong relationships. The tree structure for the Older group was significantly more complex, with deeper splits and multiple interactions between variables. This complexity reflects a methodical evaluation process in which participants in the Older group consider a broader range of variables than those in the Y-M group. The Older group utilized a total of seven features, spanning trust, experience, contextual information, perception, and demographic information, to generate sub-nodes throughout the decision tree. When comparing the two groups, notable differences arise in the variables and interaction patterns driving the decision trees. The Y-M group's decision tree was simpler and more preference-driven (with existing preferences generally maintained), whereas the Older group's tree demonstrated a more layered and complex approach to the decision-making process. These findings highlight the importance of tailoring features in building user preference models for self-driving actions to meet the unique preferences and cognitive strategies of different age groups.

Figure 74. Decision tree analysis for the Y-M group (Dependent variable: Multiple actions)



7.3.8 Discussion

Through the survey study, research questions 10 to 16 in this study are fully addressed and provide insights into user preferences with various factors and their associations. Descriptive statistics revealed higher levels of three trusts (past, current, and future expectation), the frequency of 'prefer' responses and the number of features experienced within the Y-M group compared to the Old group. In Spearman's rank correlation coefficient, the responses of 'prefer' showed statistical significance with the three trusts (past, current, and future expectation). The correlation coefficient was nearly zero between the numbers of features experienced and 'prefer' responses in the Old group. These results suggest that the Old group was less influenced by experience when choosing their preference for self-driving. Furthermore, the narrower disparity in the frequency of 'prefer' responses between high and low traffic complexity could be attributed to the Old group's preferences being more influenced by their intrinsic perceptions of self-driving. This inclination may align with a pessimistic outlook of the Old group on self-driving utilization, reflecting both their lower trust and preferences for self-driving at both high and low traffic complexity. For perceived difficulty levels, statistical significances with two traffic complexities were observed in a smaller number of situations within the Old group compared to the Y-M group. In situations where statistical significances were present, the Old group exhibited a lower error reduction probability from the Chi-square test. This suggests that perceived difficulty levels were less effective in distinguishing between two traffic complexity conditions, such as those found in Toronto and Waterloo, within the Old group compared to the Y-M group. This finding mirrors the statistical relationship observed between perceived difficulty levels and binary preferences for self-driving as analyzed through error reduction probability and proportion of variance. Consequently, it was concluded either that an additional variable would be necessary for the binary preference model for self-driving within the Old group or that the deployment of the model would need to be limited to certain situations. Regarding the action of 'being informed,' a few situations with low traffic complexity, such as 'police' and 'Turn Right' situations, were found to be less preferable. However, the preference difference between the two age groups was marginal, reaching a maximum of around 12%. Overall, this action was highly preferred in both age groups, with over 80% preference. Furthermore, the comparison between human-engaged and -disengaged actions revealed insignificant differences between the two age groups, with variances within approximately 12% in specific situations. However, across the five categories of preferred actions, correspondence analysis clearly demonstrated that age group categorization influenced situational preferences towards particular actions or neutral regions. This dynamic complicates predictions, particularly concerning situation types and traffic complexity.

7.4. Predictive Model of User Preference Using Survey Data

7.4.1 Model Selection and Optimization Process

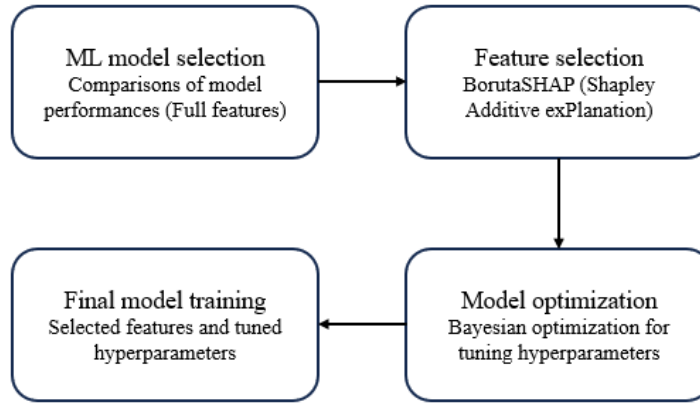
To identify high-performing ML models, classification performances for all granularity levels were tested using the Python package named Lazy Predict [290]. The package included over 30 ML models and we added three gradient-boosting ML models: Light Gradient Boosted Machine (Light GBM), XGBoost and CatBoost [291, 292, 293]. Hyperparameters for each model were optimized using the Bayesian search method based on AUC [294]. While the grid search method is commonly employed for hyperparameter tuning [295], the Bayesian search method utilized prior hyperparameter configurations to efficiently optimize hyperparameters based on AUC value. This study applied automatic searching based on the Bayesian theorem represented by the equation below.

$$x_{hyperparams} = \arg \max_{x \in X} f(x)$$

where $x_{hyperparams}$ is the set of hyperparameters that yields the highest value of the objective function. The AUC value was used as the objective function, $f(x)$, and x in the function is any value in the hyperparameter domain X defined at the beginning of the optimization [295].

In assessing model performance, other metrics such as precision, recall, F1 score, and accuracy were also compared among the selected ML models. Figure 75 illustrates our workflow for optimizing ML models. Initially, ML models were chosen based on their performance. At this stage, models that performed 10% lower than the best model in any criteria were excluded for further procedures such as feature selection and hyperparameter tuning. For feature selection, the Boruta-SHAP algorithm was employed, supporting selected features with SHAP values. Subsequently, both the grid search and Bayesian search methods were implemented to optimize hyperparameters with the selected features. In our preliminary testing of 30 ML models, the ensemble models, XGBoost and LightGBM, were the top 2 performance models. For the hyperparameter optimization, there are some hyperparameters widely used in the ensemble model optimization: maximum depth ('max_depth'), learning rate ('eta'), number of estimators ('n_estimator') and loss reduction parameter ('gamma' in XGBoost and 'min_gain_to_split' in LightGBM). By limiting the maximum depth parameter, optimized models can avoid overfitting. The learning rate and the number of estimations can be used to efficiently train ML models and possibly prevent overfitting by early stopping as well. The loss reduction factor is one of the regularization parameters used to reduce model loss between training and testing.

Figure 75. Model training procedure



In Table 31, optimized hyperparameters from granularity level 2 to 5 are presented. With the optimized feature set and hyperparameters, the selected models were trained and tested using randomly chosen 80% and 20% of the observations through cross-validation. The total number of observations was 1512 in the younger group and 1530 in the older group.

Table 31. Optimized hyperparameter ranges for user preference ML models at the four granularity levels (from the granularity level 2 to 5)

Hyperparameters	Age: 23-44	Age:60+
eta	0.218~0.334	0.097~0.388
max_depth	3.466~3.768	2.254~4.836
n_estimator	76.8~144.3	144.6~164.8
gamma	0.637~0.958	0.623~0.801

BorutaSHAP feature selection

To determine the optimal feature set, we implemented the wrapper selection method introduced in 2020, Boruta SHapley Additive exPlanations (BorutaSHAP) [296, 297]. This algorithm combines two approaches, Boruta and Shapley values. Boruta is a statistically grounded algorithm that selects a feature set by comparing importance values in permutation sets. These sets include both the original feature set and randomly shuffled samples of each feature referred to as 'shadow features.'

$$\phi_i(v) = \sum_{S \in N \setminus \{i\}} \frac{|S|! (n - |S| - 1)!}{n!} (v(S \cup \{i\}) - v(S))$$

where ϕ_i is the Shapley value of the i -th feature, N is the total number of features. S is the sum that extends over all subsets of N not containing feature i , $v(S)$ is the marginal value of the characteristic function for the features present in the coalition S , and $v(S \cup \{i\})$ is the marginal value of the characteristic function for the features present in the coalition S including feature i [298].

From the weighted averages of the shadow features, the highest value was employed as the threshold. In determining the significance of features in the original set, the weighted averages of the original features were compared against this threshold, and a feature was labelled as an ‘important feature’ if it surpassed the threshold.

7.4.2 Preference Classification: Granularity of Driver Preferences

Driver preferences for self-driving are categorized into two distinct levels [170]. The simplest level involves a binary classification determining whether a driver prefers self-driving or not ('prefer' and 'not prefer'). The highest granularity level employs five two-stage actions: 1) ‘Act-Inform’; 2) ‘Inform-Act’; 3) ‘Act-No Inform’; 4) ‘Inform-Consent’; 5) ‘Alert-Handover’ as described in Table 32. However, some studies show that people tend to have different degrees of preferences for AVs as automation level changes and the difference is also shown in age group changes [299, 300]. Therefore, we consider a design of multiple granularity levels of AV behaviours. The order and types of actions influenced driver preferences, varying based on the driving context or surrounding environment. These five categorized action groups offered a higher granularity level than binary preferences in self-driving modes. To represent various granularity levels, it is essential to decide how to characterize each action. The fundamental characterization involves whether it is a human-engaged or AV-engaged action. ‘Inform-Consent’ and ‘Alert-Handover’ are highly relevant to human-engaged actions, while the others are relevant to AV-engaged actions. At the lowest granularity level, ‘Inform-Consent’ is a crucial action representing AVs’ user-centred design since it heavily depends on the user’s choice. In terms of self-driving mode decisions, ‘Inform-Consent’ falls into a gray zone, not easily grouped as either human-engaged or AV-engaged. To simplify the lowest granularity level, it is replaced with binary preferences regarding whether users prefer self-driving modes. Moving from the lowest level to the next, users can choose between ‘Inform-Consent’ and ‘Alert-Handover’ instead of ‘AVs not preferred.’ This provides users with precise self-driving action options. Between ‘Inform-Consent’ and ‘Alert-Handover,’ users can participate in selecting self-driving modes or experience AVs forcefully handing over control in response to highly cautious traffic situations. Moreover, AVs can differentiate informing function characteristics between the two actions, e.g., informing type, intensity, frequency, etc. This allows users to personalize their AVs’ informing actions for future use. From granularity level 3 to 4, two degrees of the AV-engaged group are provided by

determining whether AVs inform users of busy areas or hazardous traffic situations regardless of informing timings. Users typically view informing actions as antecedent actions when AVs promptly detect potential risks. At granularity level 4, users have no opportunities to choose informing timings for their AVs under not-spotted traffic situations but minor levels of hazardous situations. To address this, users may opt out of certain conditions for informing actions that happen within a few seconds to avoid short notice. Another granularity level, level 5, is added to include informing timings both before and after certain traffic conditions. This offers users the chance to personalize informing timings based on traffic conditions for future use.

Table 32. Preferred actions in various granularity levels [243]

Granularity levels				Explanation of Action
Level 5	Level 4	Level 3	Level 2*	
Act-No Inform	Highly AV-engaged			No informing
Act-Inform	AV-engaged 'Informing'	AV-engaged	Preferred AV	Informing after
Inform-Act				Informing before
Inform-Consent	Inform-Consent	Inform-Consent	Gray zone	User's choice
Alert-Handover	Alert-Handover	Alert-Handover	Not prefer	Human driving

7.4.3 Classification Results of Predictive Models

In each age group, the dataset comprises over 1500 responses, surpassing the number of features by more than 60 times. This extensive dataset provides a robust foundation for constructing machine learning (ML) models to predict user preferences for self-driving behaviours. We conducted a comparative analysis of the performances (classification accuracy, F1 score, and AUC) of two ML models before and after applying the BorutaSHAP algorithm and Bayesian hyperparameter optimization. The dataset was randomly shuffled and divided into an 80/20% split for model training and testing. The top-performing ensemble models, XGBoost and LightGBM, consistently emerge as the top two models for all age groups among the top 30 ML models. Both the initial and final models achieve accuracies exceeding 80% for the lowest granularity level in Table 33. Classification accuracies increase by up to 6.8% after the model optimization process. F1 score maintains similar degrees as the classification accuracies at each granularity level for all age groups. The average AUC values are over 0.9. ML performance tends to decrease as the granularity level increases, which is attributed to the correlation between the granularity

level and the number of classes in ML prediction. The granularity level serves as a trade-off factor in adjusting prediction performance due to the complexity of the ML model in this study. In user-centred design, striking a balance is vital, offering an appropriate number of options aligned with user goals and expectations. The user-centred design should present options that users can easily comprehend and navigate to avoid overwhelming users with excessive choices for an enhanced overall user experience. Moreover, adjusting the granularity level can enhance customization in the autonomous vehicle (AV) system by allowing users to choose an appropriate level. The study emphasizes the importance of considering the number of options and the accuracy of users' preferences in two-stage behaviours.

Table 33. The comparison between the baseline model and the final model performances from the granularity level 2 to 5 (Optimized model results)

Group Age	Granularity levels	Acc.	F1 score	AUC
23-44 (Ob.=1512)	Level 2	0.858 (+0.038)	0.838 (+0.016)	0.911 (+0.025)
	Level 3	0.815 (+0.024)	0.809 (+0.019)	0.908 (+0.018)
	Level 4	0.756 (+0.036)	0.756 (+0.035)	0.893 (+0.009)
	Level 5	0.753 (+0.043)	0.749 (+0.039)	0.894 (+0.001)
60+ (Ob.=1530)	Level 2	0.850 (+0.030)	0.838 (+0.017)	0.895 (+0.003)
	Level 3	0.808 (+0.058)	0.790 (+0.035)	0.893 (+0.001)
	Level 4	0.769 (+0.068)	0.768 (+0.064)	0.888 (+0.003)
	Level 5	0.692 (+0.012)	0.693 (+0.012)	0.875 (+0.002)

- (): Performance variation post-optimization

7.4.4 Feature Importance Analysis

The evaluation of feature importance was conducted across all age groups. The summation and average of Shapley values for the top eight features exhibit small decreases or fluctuations as the granularity level increases in Table 34, with a consistent range of 11.6 to 12.4. The standard deviation (SD) of Shapley values for the top eight features decreases in the ML model of the younger group (23-44) with increasing granularity, while there is a slight decrease in the ML model of the older group (60+) as depicted in Figure 76. This suggests that with higher granularity levels, Shapley values become more distributed. The younger group experiences a more pronounced distribution of feature importance compared to the older group, indicating that ML model complexity tends to prioritize a different number of features. In terms of feature importance, Table 35 presents the ranking of the top 8 features. Across all age groups and granularity levels, the pedestrian-related feature consistently holds the top position. In the older group model, the 'Driving Experience/Age' feature secures the second rank, with either the trust or the average

preference rate feature occupying the third position. In contrast to the older group model, the younger group model exhibits variability in features ranked second and third. The older group consistently demonstrates more uniform feature priorities than the younger group model across all granularity levels, suggesting a consistent perspective in choosing preferences for self-driving behaviours. Conversely, the younger group displays different prioritization of feature types and their ranks as the granularity level changes. Figure 76 and Table 35 illustrate that the feature importance of the top 8 features is more distributed in the younger group model compared to the older group model. This indicates that the younger group is more sensitive to prioritizing feature types with changes in granularity levels. The findings imply that the older group maintains a consistent perspective in choosing preferences for self-driving behaviours across various granularity levels, while the younger group exhibits greater variability in their feature prioritization. Figure 76 illustrates the changes in importance values of the top 8 features as the granularity level increases. For the 'Driving Experience/Age' feature, the older group model consistently ranks second across all granularity levels. In comparison, the younger group model prioritizes it between the second and sixthth rank in Table 35. Simultaneously, the importance of 'Avg. Preference Rate' decreases with increasing granularity in both age groups, as depicted in Figure 76. The introduction of the 'Inform-Consent' option at granularity level 3 brings about shifts in feature importance. The younger group increases the importance of 'Driving Experience/Age' and 'Traffic Flow', prioritizing these features along with 'Pedestrian-related' and 'Avg. Preference Rate.' In contrast, the older group highly prioritizes pedestrian features with an increase in the importance of 'Vehicle Behaviour' and no change in 'Driving Experience/Age.' The preference results across this study indicate that the consent option is preferred by the younger group more than the older group. Written survey responses further highlight perceptual differences in the user consent option between the two age groups. The younger group frequently perceives informing actions and the user consent option as safer, informative and highly user-centred. On the other hand, the older group expresses a preference for immediate vehicle control, which bypasses the user consent process based on their trust in self-driving capabilities. The older group shows a much lower difference between 'Inform-Consent' and 'Alert-Handover' compared to the younger group, which has consistent differences across all situations except for low traffic volume. In conclusion, the results and written responses point to perceptual differences in the user consent option between the two age groups, emphasizing the need for tailored approaches in addressing preferences and perceptions related to self-driving behaviours. At the next granularity level, the introduction of the new option, self-driving with informing actions, becomes a preferable choice for both age groups. They express a preference for informing with a self-driving option over the two classical AV actions. Discrepancies in preferences are more pronounced in the Y-M group than the Older group, particularly in low-traffic volume situations, as previously highlighted. In terms of feature importance, 'Traffic Flow' and 'Vehicle Behaviour' are valued

similarly by both age groups. For other features, the importance differences between the two age groups remain consistent with the patterns observed at granularity level 3. In particular, ‘Traffic Unexpectedness’ has the lowest feature importance, below 4% in total. This suggests that the perspective on feature importance between both age groups has not changed significantly. Only the importance of two features in low feature ranks has become similar between granularity levels 3 and 4. In written responses, it reveals that both age groups prefer the informing action for additional safety and alertness. However, the younger group showed a larger number of responses indicating this preference than the older group. Despite these preferences, it is deemed insufficient to consider perceptual differences between the two groups as evident. Instead, there is a difference in preference rates for informing actions compared to classical AV actions. This observation aligns with the higher feature importance of the ‘average preference rate’ observed in the younger group. At the highest granularity level, the introduction of informing timing, such as pre- and post-event options, can be a new aspect of AV behaviour. This includes actions taken before or after encountering high-traffic volume areas or potential risk situations. From granularity level 4 to the highest level, there is a notable shift in feature importance in the younger group. The importance of the ‘Driving Experience/Age’ feature decreases by 40%, while other features, particularly ‘Traffic Flow’ and ‘Vehicle Behaviour,’ experience a significant increase. In contrast, the older group maintains a high value for the ‘Trust’ feature but shows a comparatively smaller increase in the importance of ‘Traffic Flow’ and ‘Vehicle Behaviour’ compared to the younger group. This suggests that the necessity of informing timing and how the timing affects drivers in AVs are perceived differently in the two age groups. In written responses to the survey, some participants in the older group view post-informing action as either an incidental feature of AVs, overcommunications or unnecessary action. However, most participants in the younger group regarded pre-informing action as an additional safety feature. This perceptual difference manifested in a divergence in preference responses for pre- and post-informing actions, especially at low traffic volumes (the difference between ‘Act-Inform’ and ‘Inform-Act’ is less than 10% at the high traffic volume but over 40% at the low traffic volume). The younger group significantly prefers pre-informing compared to post- and no-informing actions, with a response difference exceeding 40%. In contrast, the older group shows a smaller difference among pre-, post- and non-informing actions within a 10% response difference. This aligns with the general trend that younger people are more willing to adopt new technology such as ADAS, compared to older individuals [299]. Additionally, the preference for AVs tends to decline with increasing levels of automation technology in the elderly population [300]. Understanding perceptual differences and user preference variations across multiple granularity levels is crucial for designing a user-centred AV system, e.g., self-driving behaviours tailored to different age groups.

Table 34. Descriptive statistics: Shapley values of the top eight features

Granularity	Group age: 23-44			Group age: 60+		
	Sum	Avg	SD	Sum	Avg	SD
Level 2	11.8	1.48	0.90	12.2	1.53	0.82
Level 3	12.4	1.55	0.69	11.8	1.47	0.71
Level 4	11.7	1.46	0.56	11.7	1.46	0.67
Level 5	11.6	1.45	0.35	12.1	1.51	0.73

- Sum: Summation - Avg: Average - SD: Standard deviation

Table 35. Feature importance ranks of the three age groups with four granularity levels from Level 2 to Level 5 evaluated by Shapley values.

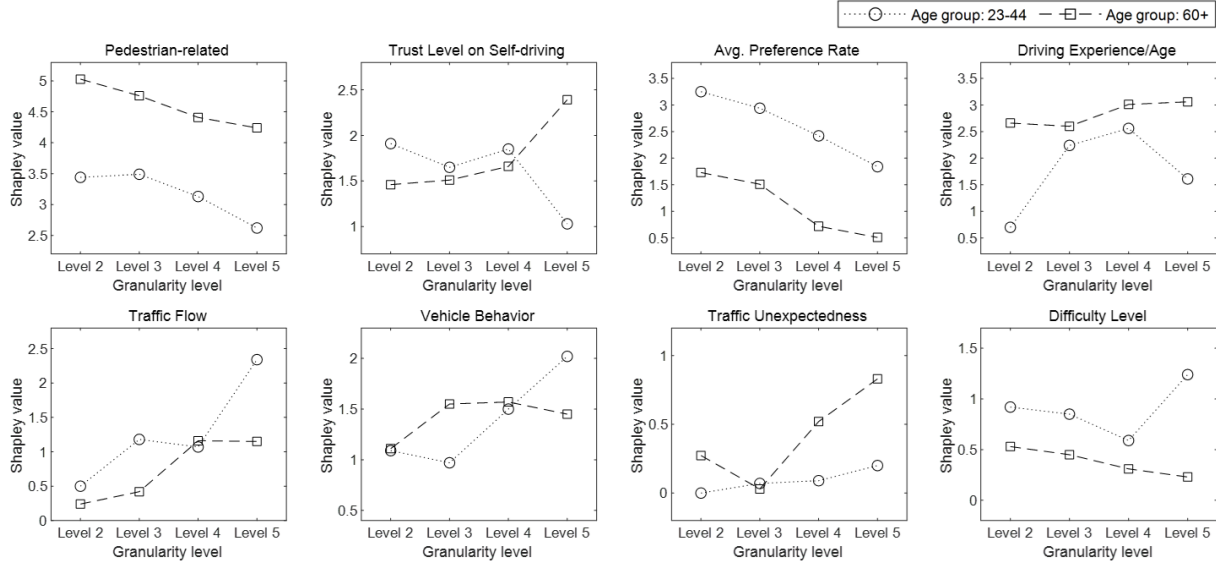
Age	Y-M: 23-44			
Granularity	Level 2	Level 3	Level 4	Level 5
Pedestrian-related	R 1 >	R 1	R 1	R 1
Trust	R 3	R 4	R 4	R 7
Avg. Preference Rate	R 2	R 2	R 3	R 4
Driving Experience/Age	R 6	R 3	R 2	R 5
Traffic Flow	R 7	R 5	R 6	R 2
Vehicle Behaviour	R 4	R 6	R 5	R 3
Traffic Unexpectedness	R 8	R 8	R 8	R 8
Difficulty Level	R 5	R 7	R 7	R 6

Age	Old: 60+			
Granularity	Level 2	Level 3	Level 4	Level 5
Pedestrian-related	R 1 >>	R 1 >>	R 1 >	R 1
Trust	R 4	R 3*	R 3*	R 3
Avg. Preference Rate	R 3	R 3*	R 6	R 7
Driving Experience/Age	R 2 >	R 2 >	R 2 >	R 2
Traffic Flow	R 8	R 7	R 5	R 5
Vehicle Behaviour	R 5	R 3*	R 3*	R 4
Traffic Unexpectedness	R 7	R 8	R 7	R 6
Difficulty Level	R 6	R 6	R 8	R 8

- R # : Rank of the feature, e.g., R 1 = 1st ranked feature

- $R \#$: The Shapley value difference between the N^{th} and the $N+1^{\text{th}}$ ranked feature is less than 0.1 (Ranked same).
- $R \# >$ and $R \# >>$: The Shapley value difference between the N^{th} and the $N+1^{\text{th}}$ ranked feature is larger than 1.0 and larger than 2.0.

Figure 76. Shapley values of the top eight features in ML models with all granularity levels.



7.4.5 Discussion

This study delved into exploring user preferences for self-driving behaviours in both age groups and the granularity levels (from 2 to 5). By leveraging a robust dataset comprising over 3000 responses, my analysis primarily focused on evaluating ML models. XGBoost and LightGBM consistently demonstrated superior performance when compared to other conventional ML models, e.g., SVM and LR. In examining the lowest granularity level representing binary preferences for self-driving modes, the trained models exhibited an impressive average accuracy of 84%. As the granularity level increased, reflecting a more detailed categorization of self-driving behaviours, the complexity of the preference model proportionally escalated. The impact of the granularity levels on ML performance emphasized the necessity of both improving preference models and choosing the appropriate number of AV behaviour options. By increasing the granularity level from 2 to 3, the younger group (23-44) highly preferred the user consent option in survey choices and written responses compared to the older group (60+). Also, the result of feature importance analysis is shown differently in both age groups. These findings revealed perceptual differences in user consent options between the age groups. From granularity levels 3 to 4, both age groups regarded the 'informing' action as an additional safety feature in written responses, providing users with alertness and chances to gain insights into informed traffic situations. The high-rank features' importance in both age groups showed no change from granularity levels 3 and 4. The younger group had a higher

preference for the informing action than the older group compared to the classical AV actions. At the highest granularity level, informing timings are added to self-driving action, both pre- and post-event. Both age groups showed significant differences in feature importance analysis. The younger group highly values ‘Traffic Flow’ and ‘Vehicle Behaviour’ features but the older group values ‘Driving Experience/Age’ and ‘Trust’ features for informing timings. In terms of preferences, the younger group strongly preferred pre-informing action compared to post-informing and no informing.

7.5. Predictive Model of User Preference Using Experiment Data

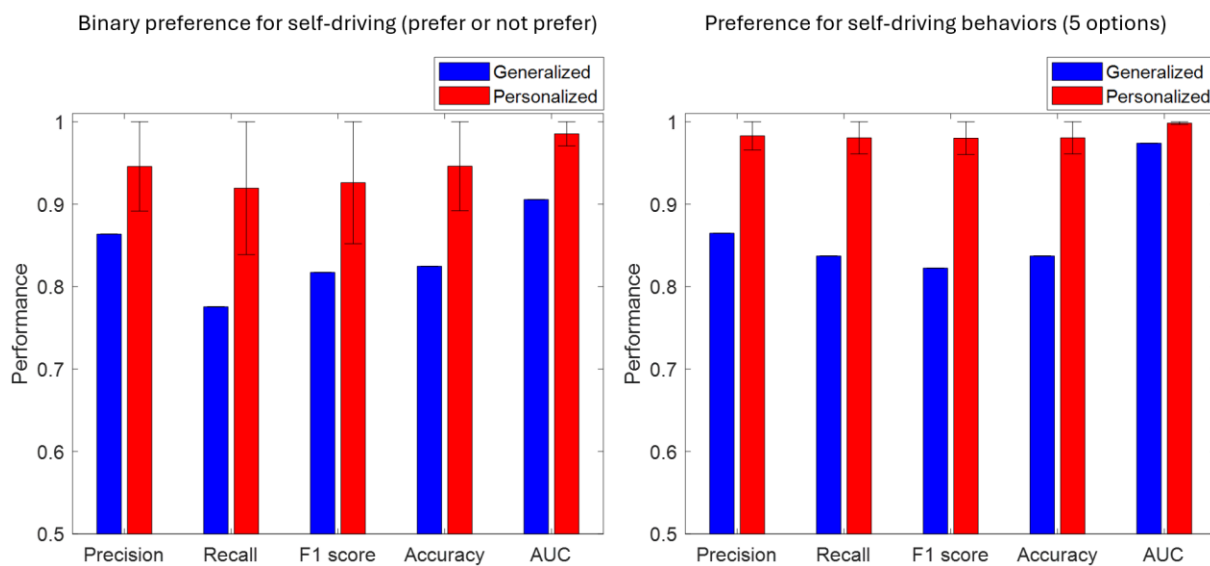
In this experiment, unlike the previous user preference study that relied on survey data, we attempted to build a user preference model by utilizing data obtained from physiological signals (EEG, BVP, GSR and Temperature) and an eye-tracking system as features. By recording these signals, it is possible to secure more feature data than from survey data, and we anticipate being able to capture intrinsic physiological responses that vary according to different traffic situations. However, it may also allow us to make more accurate inferences of preferences.

7.5.1 Performance of User Preference Models

Figure 77 illustrates the ML performance of user preference models for binary and multi-class models, respectively. Each figure shows a result of a generalized model combining data from all participants and personalized models tailored to individual participants. For both binary classification (self-driving or not) and multi-class classification (five AV behaviours), personalized models consistently outperform generalized models across all metrics: precision, recall, F1 score, accuracy and AUC. In the binary classification task, the personalized models achieved a precision of approximately 0.95 compared to the generalized model of 0.86. This indicates that personalized models are significantly better at correctly predicting user preferences without making as many false positive errors. Similarly, the recall for personalized models is around 0.92, while the generalized model shows a recall of about 0.77, suggesting that personalized models are more effective in identifying true positive cases. Consequently, the F1 score, which balances precision and recall, is higher for personalized models at 0.93 than for the generalized model at 0.80. Accuracy also shows a substantial improvement in personalized models, 0.96 over the generalized model, 0.81. The AUC reflecting the ability to distinguish between classes is nearly perfect for personalized models, with around 0.99 compared to the generalized model, which shows around 0.9. For the multi-class classification task, the performance trends are similar in all metrics. The average performance of personalized models clearly shows more than 10% higher except for AUC, which showed over 0.95 for both generalized and personalized model results. Furthermore, in multi-class classification,

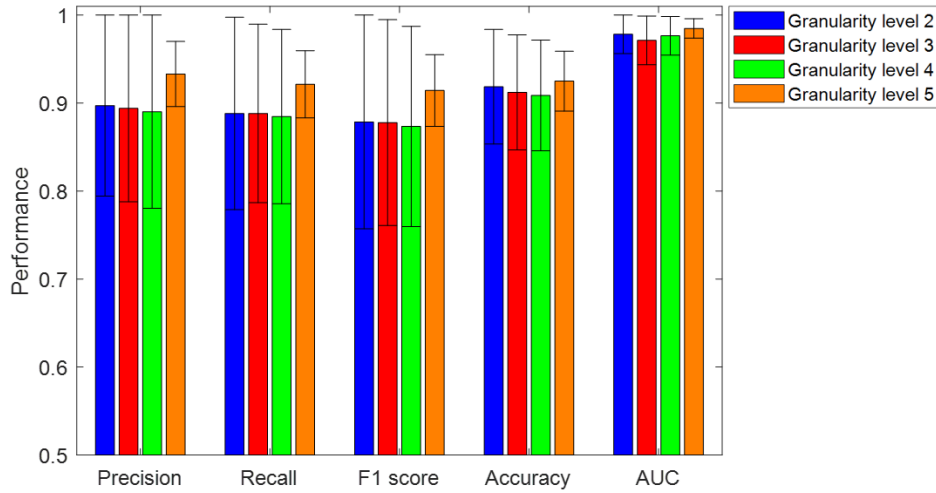
a personalized model results in much lower standard deviations of all metrics compared to the performance of the binary classification model. These results highlight the advantages of personalized ML models in capturing the unique physiological responses of individuals, leading to significantly higher performance in predicting user preferences. In contrast, generalized models fail to account for individual differences, although more straightforward to implement on a large scale, resulting in lower performance. It underscores the importance of personalized models in developing adaptive systems that can accurately predict and respond to user preferences. Also, it can enhance user experience and satisfaction.

Figure 77. Classification performance of user preference models. (Left) Two options: Prefer or not prefer self-driving. (Right) Five options: Act-Inform, Inform-Consent, Alert-Handover, etc



In Figure 78, the classification performance across granularity levels 2 to 5 shows that the model maintains consistently high performance across all metrics. There is no decline in performance as the granularity increases, suggesting that the model effectively handles the increasing complexity of additional classes. Rather, as the granularity level increases, the standard deviation decreases. This suggests that the model predictions become more consistent and stable with higher granularity levels. This result may indicate that the distribution between classes becomes clearer at higher granularity, enabling the model to distinguish them more effectively.

Figure 78. Classification performance of user preference models (Granularity level 2~5)

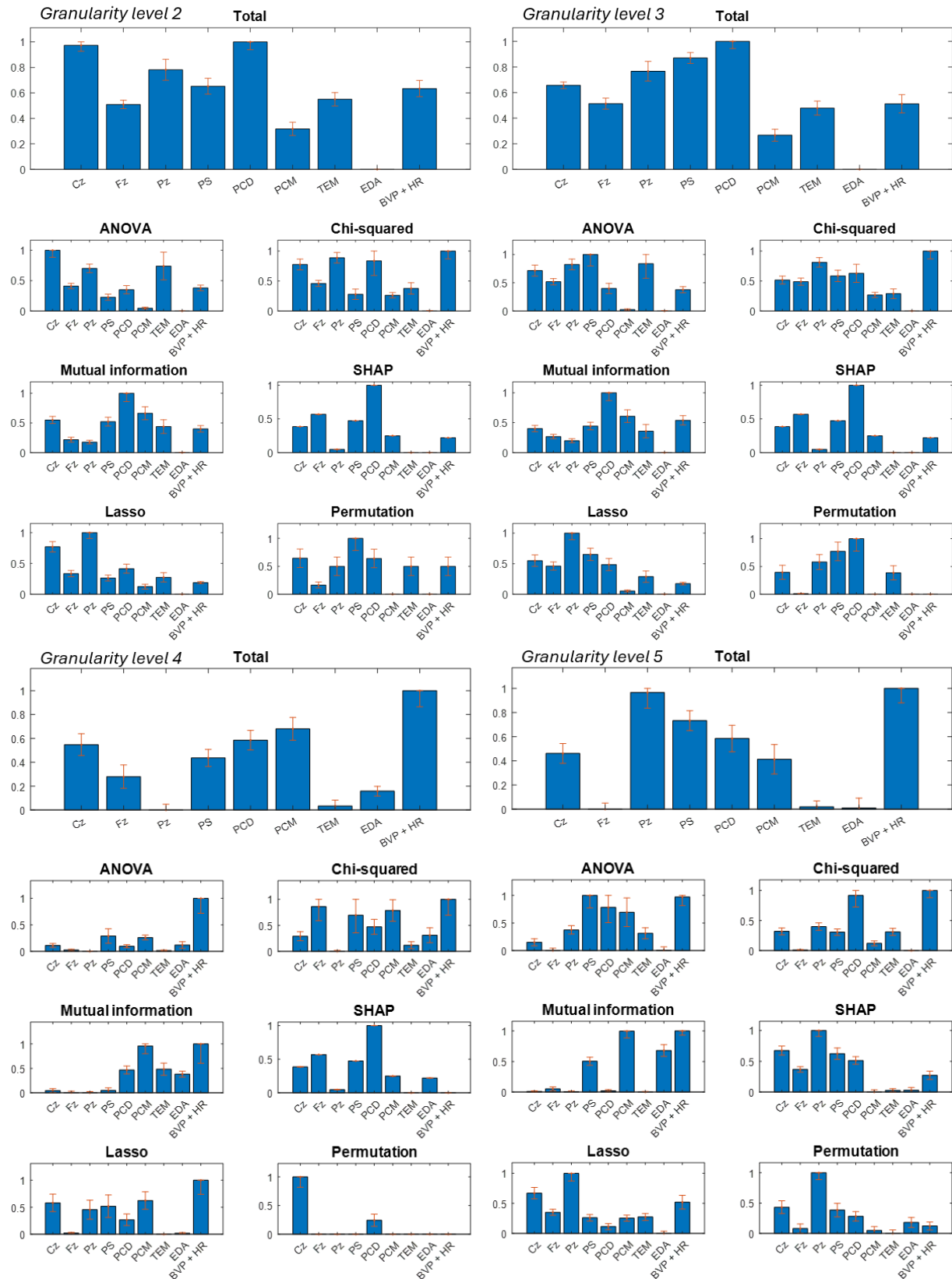


7.5.2 Feature Importance Analysis

This section provides an analysis of features used in building ML models of user preferences for autonomous vehicle (AV) behaviours. Extracted features are from EEG channels (Cz, Fz, Pz), pupil size (PS), pupil center degree changes (PCD), pupil center magnitude changes (PCM), body temperature (Tem), electrical dermal activity (EDA) and heart rates (BVP + HR). These features are analyzed across different AV behaviour classification granularity levels from binary classification (granularity level 2) to more detailed classifications (granularity level 5). In this section, we aim to understand how the importance of these features evolves as the granularity level changes with more behaviour options (and the complexity of classification also increases). The AV behaviours are categorized into four distinct granularity levels from Level 2 to Level 5 as previously shown in this study. These categories represent varying degrees of user involvement and preference for AV behaviours, ranging from minimal user intervention to traffic situations where users actively prefer to engage. The feature importance analysis employs several feature selection methods, revealing significant trends as classification complexity increases. The results of feature importance are shown in Figure 79. For the binary classification model (granularity level 2), these Cz and the PCD features exhibit considerable importance for distinguishing user preferences between self-driving and manual driving. In a more detailed analysis, the average of EEG features is higher than pupil features but not significantly different. As the classification complexity increases to three classes (granularity level 3 which includes ‘Inform-consent’), visual attention and scanning to perceive traffic situations can be reflected in the increased importance of pupil features. This suggests that in terms of the model interpretability related to neural activity, there is a relative shift towards

greater sensitivity to changes in pupil size. This is likely because the user's judgment becomes more critical with the addition of the 'Inform-Consent' option depending on surrounding traffic conditions. This can be explained by the dynamics and patterns of PS and PCD reflecting salient events during the user's perception of the surrounding traffic environment and making decisions. Therefore, the introduction of the 'Inform-Consent' category indicates the need for more detailed eye-tracking data to understand user consent behaviours. At the granularity level 4, where informing actions with self-driving is included as an option, there is a significant decrease in the overall scale of feature importance except for the BVP+HR feature. Particularly, while EEG or pupil features were the most influential in classification models at granularity levels 2 and 3, at granularity level 4, the BVP+HR feature shows the highest importance without any other features displaying similar significance. This reduction in overall feature importance is likely due to the increased complexity of the model and the diversity of user preferences brought about by the inclusion of the informing action with self-driving. Interestingly, at granularity level 5, the BVP+HR feature retains its position as the most important feature, consistent with its significance at the granularity level 4. At granularity levels 4 and 5, the addition of more options for AV-engaged behaviours shows a strong correlation with the cardiac activity feature, BVP + HR. However, at granularity level 5, the importance of EEG and pupil features, which decreased at the granularity level 4, begins to rise again. Despite this increase, their importance remains lower than at the granularity levels 2 and 3, except for the Pz feature. The Pz and BVP+HR features emerge as the two most significant features. In summary, adding new AV behaviour options at each granularity level significantly impacts feature importance. For instance, the transition from 'AV-engaged' to 'Inform-Consent' and 'Alert-Handover' introduces situations where users must process and respond to information. This change can shift the importance of all features, reflecting cognitive loads from decision-making based on salient information. These findings suggest that as AV behaviour options become more detailed, users can rely more on physiological responses or rely on other data to make informed decisions, e.g., traffic environment perceptions introduced in the next chapter.

Figure 79. Comparison of feature analysis from granularity level 2 to 5



7.5.3 Discussion

The study demonstrates that personalized ML models consistently outperform generalized models across all evaluation metrics (precision, recall, F1 score, accuracy and AUC). In both binary and multi-class classification tasks, personalized models achieved higher precision and recall, indicating their effectiveness in correctly predicting user preferences while minimizing false positives and false negatives. The substantial improvement in the F1 score and accuracy further emphasizes the value of personalized models in capturing the unique physiological responses of individuals. The consistency in performance across different granularity levels, ranging from binary classification to more detailed multi-class classification, suggests that these models are robust in handling increasing complexity. Interestingly as the granularity level increases, the standard deviation in model performance decreases, indicating that the models' predictions become more consistent and stable. This trend implies that the models can better distinguish between classes at higher granularity levels, possibly due to the more apparent distribution of preferences as the classification becomes more specific. These results underscore the importance of developing adaptive systems that can accurately predict and respond to individual user preferences. The ability of personalized models to outperform generalized models highlights the need for tailoring autonomous vehicle systems to individual users, thereby enhancing user experience and satisfaction. By accounting for individual differences, these systems can provide more accurate predictions and offer a more personalized and user-centred experience. The feature importance analysis reveals the changing significance of various physiological signals as the complexity of classification increases. The importance of both EEG signals and pupil features (PS, PCD and PCM) shows high feature importance in determining binary preferences for self-driving, prefer or not prefer, at the lowest granularity level. At the granularity level 3, as the user consent option is included in AV behaviours, EEG feature importance is slightly decreased, and the importance of pupil features rises, reflecting an increased reliance on pupil movements/size change for decision-making in more complex AV behaviour options. At the granularity level 4, the importance of features from Cz and Fz is evident, whereas at level 5, Cz and Pz features become more prominent. Additionally, the importance of EEG features at level 5 increases compared to level 4, highlighting their growing role in more intricate classification tasks as the order of informing action is added to the behaviour options. Heart activity feature (BVP + HR) shows moderate importance at lower granularity levels 2~3. However, its impact on classification is significantly increased at the higher granularity level 4~5. Furthermore, as the granularity level increases, the importance of the Temp feature gradually decreases, becoming nearly negligible by level 4. The EDA feature consistently shows almost zero importance across all granularity levels.

7.6 Conclusion

Research questions regarding building user preference models for self-driving actions by using two different studies, survey-based and experiment-based studies, are fully addressed (RQ17~20). Findings through this research highlight the necessity of developing personalized and adaptive autonomous vehicle systems that account for individual differences and varying levels of self-driving behaviour granularities. By leveraging personalized ML models and understanding the shifting importance of physiological features across different granularity levels, self-driving behaviours can be better tailored to meet users' specific needs and preferences. This approach enhances user experience and satisfaction and fosters greater trust and acceptance of autonomous driving technologies. Furthermore, it allows users to design their own semi-AV systems that is able to adjust automation levels depending on traffic situations and individual preferences. The findings from this research contribute to a new paradigm of developing semi-AVs and advanced user-centred AV systems capable of delivering more accurate, reliable and personalized experiences in an increasingly complex driving environment.

Chapter 8. Conclusion

8.1 Contributions

The findings presented in Chapters 4 through 7 reveal the potential of AI-driven user perception models using an integration of physiological signals and methodologies for improving human interaction with advanced vehicle environments. The integration of physiological signals, particularly EEG and eye-tracking, as explored in Chapter 4, lays the foundation for robust artifact removal processes, and preprocessing for other physiological signals is also explained in the chapter. These preprocessing steps enable us to extract features related to cognitive and physiological states with cleaned data. It is crucial for the subsequent development of AI-driven user perception models using physiological signals.

In Chapter 5, using extracted features from physiological signals, AI-driven risk detection models are built across various traffic conditions, and the models show robust classification capabilities even as the complexity of the models increases with more classes. Risk detection models in different levels of model complexity show high performance across classification performance metrics like precision, recall, F1 scores and accuracy frequently exceeding 0.9. In particular, our models showed greater stability and higher accuracy in scenarios involving pedestrian collision risks compared to those involving vehicles, suggesting that pedestrian-related factors may present clearer patterns for the models to detect. However, regardless of collision objects between pedestrians and vehicles, the binary error classification model distinguishing between high, low potential risk, and no collisions shows high performance. Feature importance analysis revealed that the relevance of specific EEG and pupil features varied depending on the scenario. For example, EEG channels like Pz and Fz were critical in different contexts, and pupil features such as PS and PCD showed varying importance across different environments. These findings highlight the need to consider the context-specific relevance of features when developing classification models. In Chapter 6, the findings demonstrate the significant advantages of combining multiple machine learning approaches, e.g., supervised, unsupervised and semi-supervised methods, to develop robust models for predicting mental workload using physiological signals. This approach can be used to develop workload prediction models to estimate relative workloads and personalize the models. It can also be strengthened by particularly being validated against established benchmarks like the NASA-TLX. In supervised learning, the models demonstrated predictive solid power and high R^2 values, suggesting that these models are well-suited for predicting workload over time, affirming the linear relationship between experiment proceeding time and workload levels based on NASA-TLX in this study. The exploration of unsupervised methods such as K-means and Hierarchical Clustering provided more profound insights into workload levels closely tied to the number of classes. Hierarchical clustering consistently outperformed

K-means, indicating its robustness in revealing underlying patterns in complex data. The study also suggested the optimal numbers of workload prediction models as fewer clusters provided more stable and interpretable results. Simplifying the classification process can enhance the utility of classification models. In our semi-supervised learning approach, each cluster data from the unsupervised learning can be classified as one of the workload levels (classes) based on NASA-TLX results, the linear relationship between workload levels and time and distances to each cluster measured by unsupervised learning. This approach proved especially beneficial in some scenarios where labelled data is scarce when users are not highly fatigued, showing obvious symptoms when it requires estimation of relative workloads from the beginning. Furthermore, this approach is able to adjust the number of classes for workload levels, suggesting the potential of adjusting granularity in workload levels according to user needs. In the feature importance analysis, EEG signals, especially from the Pz channel, emerged as significant for binary classifications of workload, indicating a strong correlation between these signals and cognitive load. In multi-class classification tasks, however, the importance of features was more evenly distributed across EEG channels, suggesting a more complex interplay of cognitive processes. Additionally, pupil features, particularly pupil center degree and magnitude, showed substantial significance in more complex classifications, surpassing the traditionally emphasized pupil size feature. Conversely, features related to electrical dermal activity and body temperature were consistently less influential, indicating their limited role in this context. Overall, these findings highlight the importance of a multifaceted approach that leverages the strengths of different learning methods and carefully selected features to build accurate and reliable models for workload prediction.

In Chapter 7, the findings provide insights into user preferences for self-driving behaviours across different age groups, various traffic situations and levels of granularity in self-driving behaviours. As a preliminary study for building user preference models, our statistical analysis of survey data indicated that younger participants generally exhibited higher levels of trust and a stronger preference for self-driving options. The older group's preferences were less influenced by prior experience, indicating a more intrinsic skepticism toward self-driving technology. This was particularly evident in their responses to traffic complexity, where perceived difficulty levels had less of an impact on their preferences than those of the younger group. In terms of traffic situations, each traffic situation can have a strong correlation with a specific type of self-driving behaviour, e.g., informing action in 'aroundCar' scenario in Waterloo. These age, traffic complexity, and situation factors are evident in need for age-specific, scenario-specific and complexity-considered models of user preference for self-driving behaviours. Furthermore, the importance of other factors, such as surrounding vehicle behaviours, traffic flow, etc., is highlighted in certain granularity levels. These insights emphasize the need for adaptive AV systems that are able to provide customizable options in self-driving behaviours that cater to diverse user

preferences. The analysis based on physiological data pointed out that user preferences can be modelled by features extracted from physiological signals that represent user intrinsic responses to various factors in each traffic situation. The preference prediction results reinforced the value of personalized models that consistently outperformed generalized models across ML performance metrics. Based on personalized preference models, there were no noticeable changes in the model performances as the granularity level increased, indicating that personalized models provided more consistent and accurate predictions. This highlighted that individual physiological responses could be used for adaptive and personalized systems that can respond to individual needs in real-time. It can possibly enhance user satisfaction and trust in autonomous vehicle technologies.

8.2 Conceptual Integration of User Perception Models

This thesis aimed to develop user perception models within the context of advanced vehicle technology, focusing on creating an integrated perception model rather than relying on a single model. For more details of the integrated perception model, Figure 80 presents a schematic diagram demonstrating the flow from physiological signal preprocessing to the final perception models. Feature extraction under the appropriate epoch conditions is essential for each perception model, as these features are then directed to their respective models. The integrated user perception model marks a significant advancement in human-vehicle interaction, particularly for semi-autonomous vehicles (semi-AVs). This holistic approach has the potential to revolutionize advanced vehicle systems in a variety of ways.

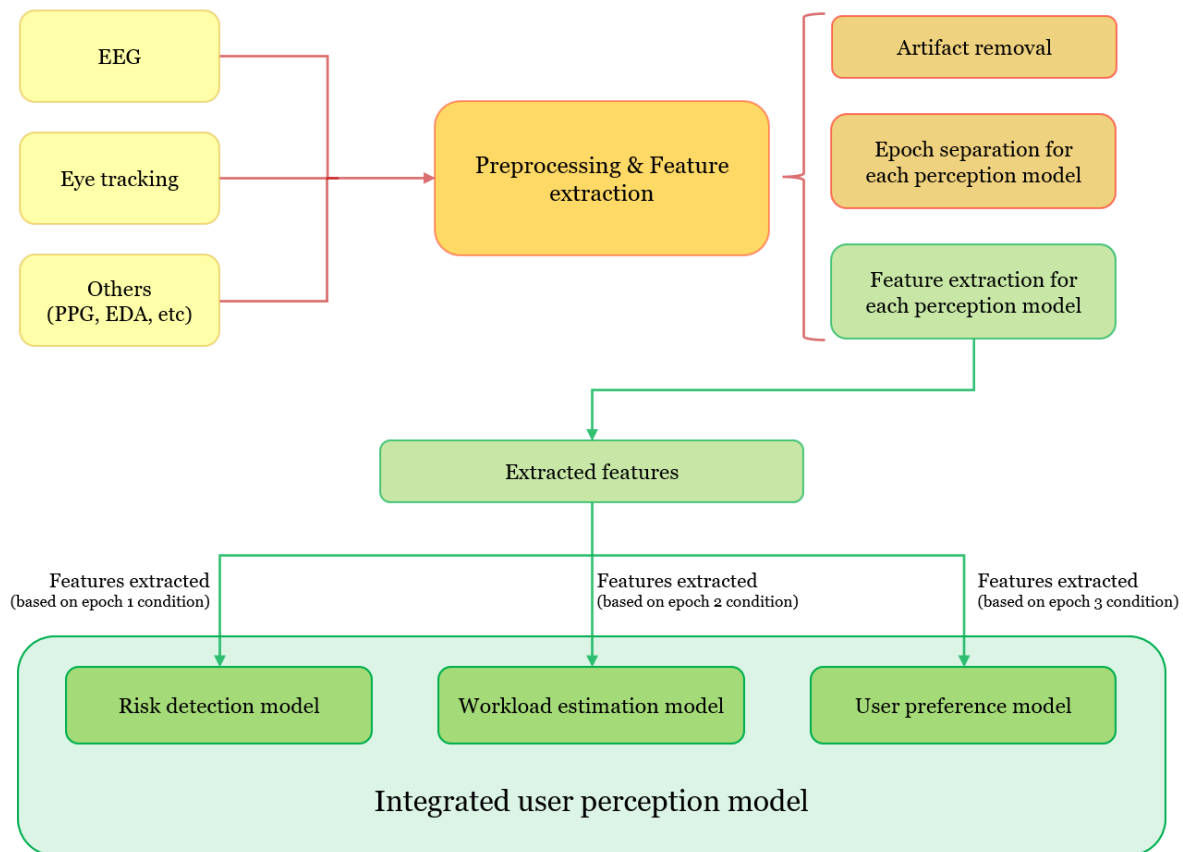
1) Enhanced Safety and Responsiveness

- **Real-time Adaptation:** The integration of risk detection, workload estimation and user preferences allows the semi-AV system to adapt in real time to the user's physiological and cognitive states. For instance, if the system detects elevated stress levels (high workload) coupled with an imminent risk (as detected by the risk model), it can adjust driving behaviours or alert the user accordingly, thereby enhancing overall safety.
- **Proactive Risk Management:** By integrating these models, the vehicle can anticipate potential risks by considering both environmental factors and the user's current state rather than merely reacting to risks as they occur. This proactive approach has the potential to significantly reduce the likelihood of accidents, especially in complex and dynamic driving environments. The features derived from environmental factors and the user's current state at the anticipated risks can be utilized to select the most appropriate behaviour from either the user preference model in this thesis or a new behaviour model for semi-autonomous vehicles.

2) Personalized User Experience

- **Tailored Driving Behaviours:** With the integration of user preference models, semi-AVs can offer a driving experience customized to individual users. This means that the vehicle can adjust its driving style based on the preferences of the user, whether they prefer a more cautious or aggressive approach, adapting in real-time based on their physiological feedback.
- **Increased User Trust and Comfort:** Personalization through the integrated model fosters a sense of trust and comfort among users. When users see that the vehicle responds to their specific needs and preferences, their confidence in the technology will likely increase, leading to higher adoption rates of semi-AV systems.

Figure 80. A schematic diagram of the integrated user perception model



While the integrated user perception model shows significant promise, there are still areas that require further refinement. To develop a more effective model suitable for integration into advanced vehicles,

the following sections, specifically ‘8.3 Future Work,’ outline key areas for enhancement and potential improvements.

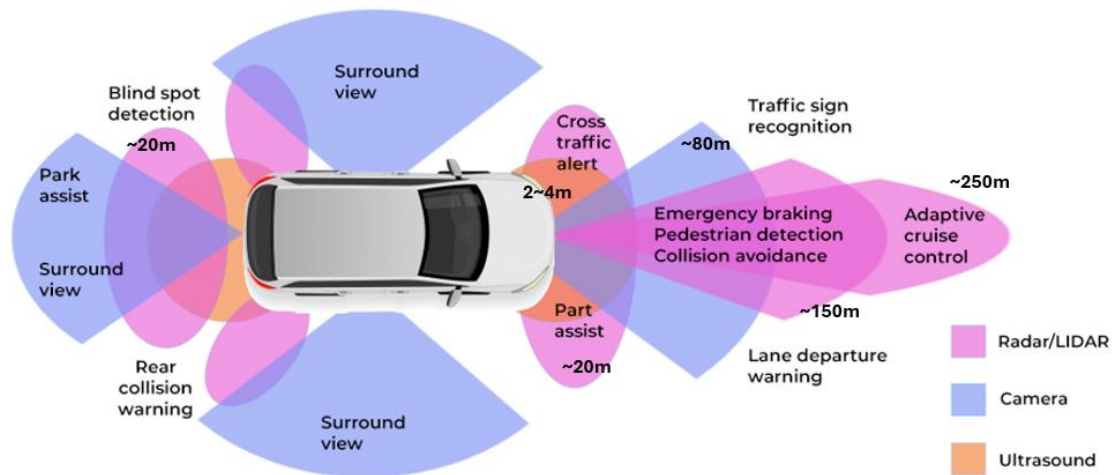
8.3 Future Work

8.3.1 Suggestions for Further Studies

1) Environment perception for surrounding traffic environment/situation

The most essential step to improve our integrated user perception model, as shown in Figure 82, is to incorporate an environmental perception dimension. Environment perception utilizes various sensors to detect surrounding traffic situations. These sensors, such as camera systems (monocular, stereo, infrared), radar (short-range and long-range), Light Detection and Ranging (LiDAR), ultrasonic sensors and GPS, are widely used in ADAS and autonomous driving systems to gather traffic data. Figure 81 demonstrates the application of these sensors in ADAS functions. Additionally, Table 35 provides a summary of environment perception elements that can be identified from the sensor data.

Figure 81. How ADAS works [301]



By using the traffic information listed in Table 36, a 3D traffic environment can be mapped in real-time. Additionally, traffic environment features derived from surrounding traffic information can be integrated with the user perception model proposed in this thesis. For example, user preferences regarding vehicle behaviours in specific traffic situations, vehicle behaviours in high-risk scenarios for enhanced safety, or vehicle behaviours adjusted according to user workload can be incorporated. These elements can be integrated with the user perception models presented in this thesis. Moreover, extensive GPS data can be

utilized to include real-time traffic volumes and various traffic events at geolocations along the upcoming itinerary or route path.

Table 36. List of surrounding and upcoming traffic environment

Function	Perceived traffic environment information
Object Recognition	Identifies objects such as pedestrians, traffic signs, traffic lights, road signs, static and dynamic obstacles, etc.
Vehicle Speed and Distance	Detects the position and movement of surrounding vehicles including the speed and distance.
Lane Detection	Detects lane markings and boundaries.
GPS	Provides geolocation, time information, current traffic volumes and any traffic events, e.g., accidents, construction, etc.

2) Expansion from environment perception to user perception model

In the user perception model, traffic environment perception can be utilized in several ways. Firstly, it can be used to identify similarities between existing traffic situations and new traffic situations, allowing for labelling and classification into the most similar types of traffic situations. By labelling new traffic environments or situations based on this traffic information, it becomes possible to predict user preferences as well as highly probable situations that may frequently trigger risk detection and increased workload without the need for additional physiological signal recordings. The sections below outline how environment perception can enhance the user preference model.

- (1) Preference models can set default preferences for self-driving behaviours at GPS-based locations with environment perception features where/when users primarily operate self-driving vehicles, e.g., peak hours, parking lots, school zones with certain time ranges, etc.
- (2) Preference models with environmental perception data can work when other features from a survey and physiological signals are not available.
- (3) Preference models can speed up the building of user preference models, e.g., analogous traffic situations/conditions in preference models from either survey data or associations between survey data and other components that can be expanded by the existing preference model.

3) Driver monitoring system

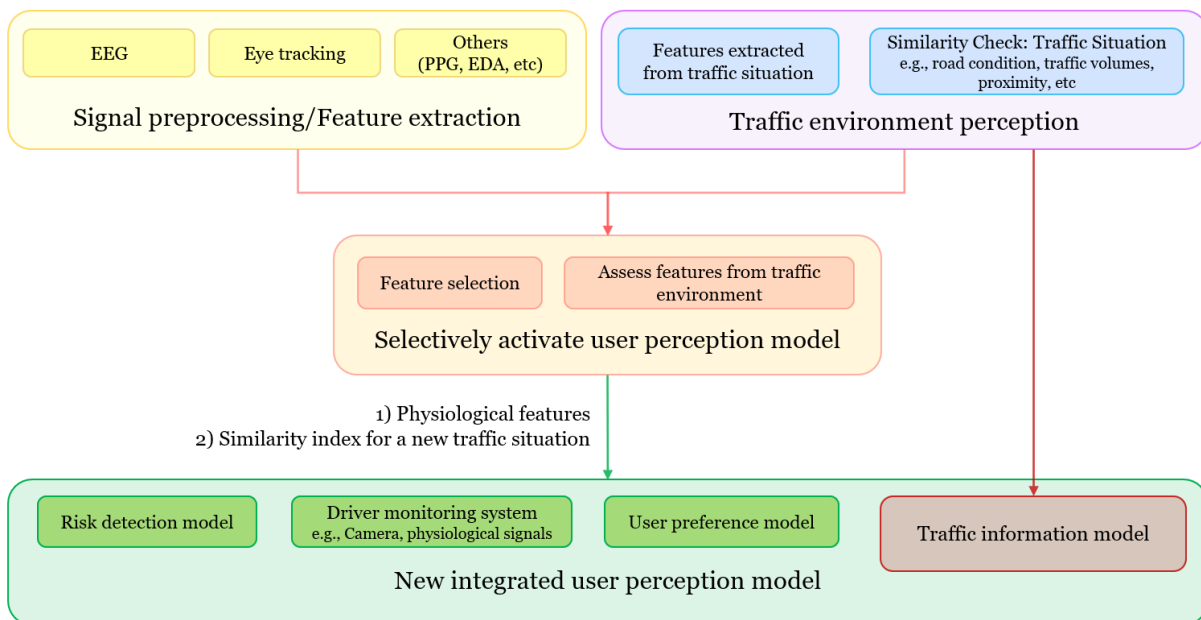
This thesis includes only a driver monitoring model focused on workload estimation. It does not encompass a detection model for clear signs of driver fatigue, such as frequent yawning, heavy eyelids and head nodding, as listed in Table 37. Additionally, certain factors, like lane drifting, slow reaction times

and missing traffic signs, can indicate levels of driver fatigue. These factors can be assessed by analyzing driving performance features and traffic information. Rather than focusing solely on workload estimation, a comprehensive driver monitoring system is presented in Figure 82.

Table 37. Detection of fatigued states of drivers

Required info.	Symptoms	Symptom explanation
Camera monitoring	Frequent yawning	Constant yawning is a clear sign of fatigue and tiredness
	Heavy eyelids	Keeping eyes open, frequent blinking, or drooping eyelids
	Head nodding	Struggling to keep their head up, frequently nodding off
Driving monitoring and traffic information	Lane drifting	Drifting into other lanes (difficulty to stay within the lane)
	Slow reaction times	Delayed responses (e.g., slow braking or steering, etc)
	Difficulty focusing	Trouble focusing on the road, traffic signs, etc.

Figure 82. New integrated user perception model with traffic information

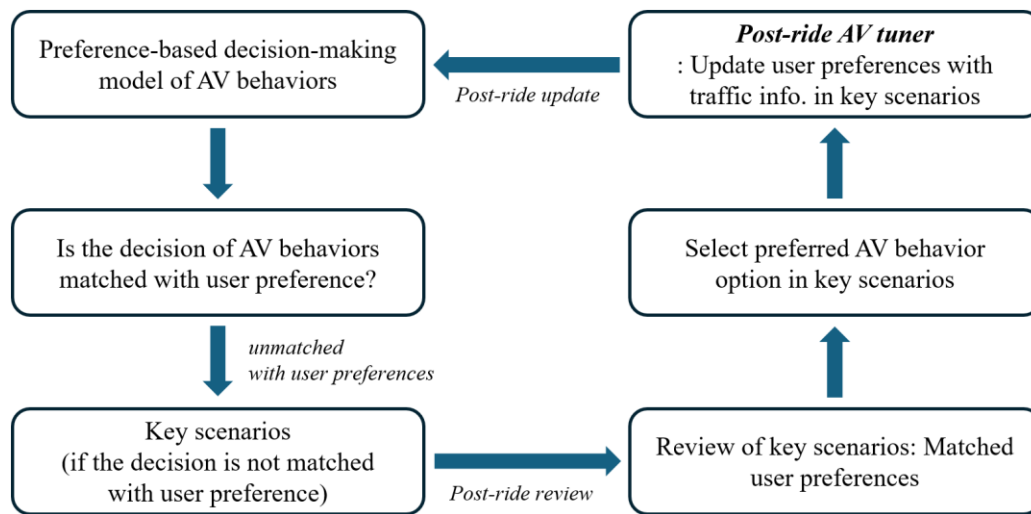


4) Post-ride tuner for semi-AV behaviours

The new integrated user perception model also requires correcting and updating false alarms or misses. Additionally, when the similarity index of a traffic situation is ambiguous, it is necessary to update the

model for that specific situation. To address these issues, functionality is needed for users to update the model themselves. This update should occur after the ride, allowing users to review and update the model by analyzing video or image data captured by the vehicle's camera for traffic situations where false alarms or misses occurred. This post-ride tuner approach can also be applied to the driver monitoring system. Figure 83 below demonstrates an example of applying the post-ride tuner approach to user preferences. When a decision from the model does not align with user preferences in specific traffic scenarios, those scenarios are categorized as key scenarios. After the ride, users can review these key scenarios and update their preferences for AV behaviours in the decision-making model. Post-ride tuning can significantly enhance the user preference model.

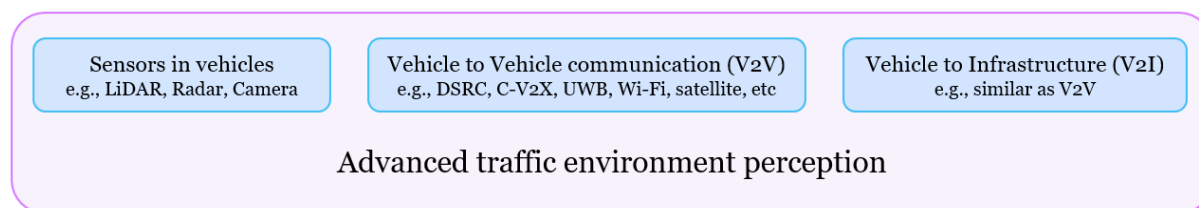
Figure 83. Post-ride AV tuner to update user preferences on the decision-making model of AV behaviours.



5) Vehicle to Everything: vehicle (V2V) and infrastructure (V2I) communications

Beyond environmental perception, there are other technological approaches to extract features regarding other vehicles and upcoming traffic conditions, such as vehicle-to-vehicle (V2V) and vehicle-to-infrastructure (V2I) communications. These technologies enable vehicles to perceive real-time traffic information, road condition monitoring and other vehicles' intentions, such as lane change, overtaking, emergency vehicles, vehicle trajectories, etc as shown in Figure 84. This information via V2V and V2I technologies can significantly enhance road safety, response time and perception of the traffic environment.

Figure 84. Advanced traffic environment perception through Vehicle-to-Everything communications



6) Other user perception models

A broader range of user perception models can be explored as future studies to enhance the adaptability and safety of advanced vehicle systems. One of the potential user perception models is the Situational Awareness Perception Model, which focuses on understanding the user's awareness of the environment and the vehicle's state to ensure informed decision-making and enhance safety. Additionally, the emotional state perception could be developed to dynamically adjust vehicle responses based on the user's emotional state, e.g., stress or anxiety, thereby improving comfort and trust in advanced vehicle systems. User attention and focus perception could be monitored for the user's attentiveness, providing timely alerts or interventions if the user becomes distracted. Eye-tracking and traffic environment perceptions could be used to develop user attention/focus models. Lastly, as a broader concept, the intention and decision-making perception for various functions and vehicle operations could be predicted by analyzing past behaviours and contextual cues which are similar to user preference models. However, this enables advanced vehicles to predict user intentions promptly. Incorporating these models would provide a more comprehensive approach to understanding and responding to diverse human factors, ultimately leading to more user-centred and responsive autonomous driving systems.

7) Other interactions between a vehicle and a user

Other interactions, such as sound/voice and gestures, could also be explored to enhance user perception and interaction within advanced vehicles. Voice and gesture controls offer a natural and intuitive way for drivers to communicate with vehicle systems, providing a hands-free and eyes-free experience that can improve safety and convenience. For instance, voice commands can be used to adjust driving behaviours, such as switching between autonomous and manual modes. At the same time, gestures can seamlessly control in-cabin features like music volume, temperature settings, or navigation inputs. The interplay between voice and gestures allows for more dynamic interactions, enabling users to switch between modalities based on context or preference. Moreover, these interaction models could be adapted to accommodate different user profiles and contexts, enhancing personalization and comfort. By integrating multimodal interaction methods such as voice and gestures with AI-driven user perception models, future

systems could more accurately interpret and respond to user intentions, leading to a more engaging and user-centred driving experience.

8.3.2 Long-Term vision

In an era where technology rapidly advances, there is an immense opportunity to rethink how users interact with intelligent vehicles, such as semi-autonomous vehicles (semi-AVs), by placing the user at the core of this evolution. My personal goal is to design various user-centred functions in intelligent vehicles that can maximize interactions between the vehicle and its users, transforming vehicles from mere tools of transportation into intelligent partners that understand, adapt and respond to human needs. By incorporating real-time user feedback, physiological data and contextual information into the vehicle's decision-making processes, the user-centred design approach based on user perceptions will keep the user central to the driving experience and vehicle functionalities. Integration of user perception into vehicles will result in vehicles that are more intuitive and responsive, fostering trust and collaboration between humans and vehicles. This dynamic adaptability to individual needs and preferences will create a more personalized and safe driving environment. This approach marks the beginning of a new era of human-AI collaboration. The future of semi-AVs is not just about automation; it is about communication between human users/drivers and intelligent vehicles that work seamlessly together. These vehicles can perceive, predict and respond to users' physiological and cognitive states, adapting their behaviours in real time. Such collaboration will enhance safety by anticipating risks, improve efficiency by optimizing vehicle behaviours based on driver comfort and readiness and provide a more engaging and enjoyable driving experience tailored to each user. My vision is to continue advancing these aspects, ensuring they become more accurate, adaptive and inclusive of diverse user needs and contexts.

References

- [1] Stanton, N. A., & Salmon, P. M. (2009). Human error taxonomies applied to driving: A generic driver error taxonomy and its implications for intelligent transport systems. *Safety Science*, 47(2), 227-237.
- [2] Shahriari, M., Örtengren, R., Aliandrina, D., & Feng, Y. (2008). Human Errors in Decision Making-Analysis of two historical cases. *Journal of Ergonomia IJE&HF*, 30(1), 47-62.
- [3] Eboli, L., Mazzulla, G., & Pungillo, G. (2017). Measuring the driver's perception error in the traffic accident risk evaluation. *IET intelligent transport systems*, 11(10), 659-666.
- [4] Öhman, S. (2017). Previous experiences and risk perception: The role of transference. *Journal of Education, Society and Behavioural Science*, 23(1), 1-10.
- [5] Nevelsteen, K., Steenberghen, T., Van Rompaey, A., & Uyttersprot, L. (2012). Controlling factors of the parental safety perception on children's travel mode choice. *Accident Analysis & Prevention*, 45, 39-49.
- [6] Christensen, J. S., Wild, H., Kenzie, E. S., Wakeland, W., Budding, D., & Lillas, C. (2020). Diverse autonomic nervous system stress response patterns in childhood sensory modulation. *Frontiers in Integrative Neuroscience*, 14, 508759. <https://doi.org/10.3389/fnint.2020.00006>
- [7] Dusi, V. (2020). Nociception, sympathetic nervous system, and inflammation: When the heart and the skin speak the same language. In *Sympathetic Nervous System* (pp. 7). Springer. https://doi.org/10.1007/978-3-030-28008-6_63
- [8] Tobaldini, E., Colombo, G., Porta, A., & Montano, N. (2019). The autonomic nervous system. In *Stress Challenges and Immunity in Space: From Mechanisms to Monitoring and Preventive Strategies* (pp. 123–143). Springer. https://doi.org/10.1007/978-3-030-16996-1_8
- [9] Gray, K., Arnott-Hill, E., & Benson, O. (2020). *Introduction to Psychology*. College of DuPage Digital Press.
- [10] Mathôt, S. (2018). Pupillometry: Psychology, physiology, and function. *Journal of Cognition*, 1(1), 1–23. <https://doi.org/10.5334/joc.18>
- [11] Kim, S. K., Kirchner, E. A., Stefes, A., & Kirchner, F. (2017). Intrinsic interactive reinforcement learning – Using error-related potentials for real-world human-robot interaction. *Scientific Reports*, 7(1), 1–16. <https://doi.org/10.1038/s41598-017-17682-7>
- [12] Spüler, M., & Niethammer, C. (2015). Error-related potentials during continuous feedback: Using EEG to detect errors of different type and severity. *Frontiers in Human Neuroscience*, 9(MAR), 123266. <https://doi.org/10.3389/fnhum.2015.00155>
- [13] Murphy, P. R., Van Moort, M. L., & Nieuwenhuis, S. (2016). The pupillary orienting response predicts adaptive behavioural adjustment after errors. *PLoS ONE*, 11(3), e0151763. <https://doi.org/10.1371/journal.pone.0151763>

- [14] Roquet, D., & Conti, F. (2021). Disentangling the association between the insula and the autonomic nervous system. *Journal of Neuroscience*, 41(14), 3051–3053. <https://doi.org/10.1523/JNEUROSCI.2225-20.2021>.
- [15] Boubchir, L., Touati, Y., Daachi, B., & Cherif, A. A. (2015). EEG error potentials detection and classification using time-frequency features for robot reinforcement learning. In *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBS, 2015-November, pp. 1761–1764)*. IEEE. <https://doi.org/10.1109/EMBC.2015.7318719>.
- [16] Critchley, H. D., Tang, J., Glaser, D. butterworth, B., & Dolan, R. J. (2005). Anterior cingulate activity during error and autonomic response. *NeuroImage*, 27(4), 885–895. <https://doi.org/10.1016/j.neuroimage.2005.05.047>.
- [17] Botvinick, M. M., Carter, C. S., Braver, T. S., Barch, D. M., & Cohen, J. D. (2001). Conflict monitoring and cognitive control. *Psychological Review*, 108(3), 624–652. <https://doi.org/10.1037/0033-295X.108.3.624>
- [18] Brázdil, M., Roman, R., Falkenstein, M., Daniel, P., Jurák, P., & Rektor, I. (2002). Error processing - Evidence from intracerebral ERP recordings. *Experimental Brain Research*, 146(4), 460–466. <https://doi.org/10.1007/s00221-002-1201-y>
- [19] Hajcak, G., McDonald, N., & Simons, R. F. (2003). To err is autonomic: Error-related brain potentials, ANS activity, and post-error compensatory behaviour. *Psychophysiology*, 40(6), 895–903. <https://doi.org/10.1111/1469-8986.00107>
- [20] Cheyne, J. A., Solman, G. J., Carriere, J. S., & Smilek, D. (2009). Anatomy of an error: A bidirectional state model of task engagement/disengagement and attention-related errors. *Cognition*, 111(1), 98-113.
- [21] Veen, F. M., van Bochove, M., Brass, M., & Notebaert, W. (2004). Cardiac and electro-cortical responses to performance feedback reflect different aspects of feedback processing. *Current Opinions on Performance Monitoring*, 15, 54-67.
- [22] Fiehler, K., Ullsperger, M., & Von Cramon, D. Y. (2005). Electrophysiological correlates of error correction. *Psychophysiology*, 42(1), 72–82. <https://doi.org/10.1111/j.1469-8986.2005.00265.x>
- [23] Wessel, J. R., Danielmeier, C., & Ullsperger, M. (2011). Error awareness revisited: Accumulation of multimodal evidence from central and autonomic nervous systems. *Journal of Cognitive Neuroscience*, 23(10), 3021–3036. <https://doi.org/10.1162/jocn.2011.21635>
- [24] Bruschi, V., Dourou, N., Iadarola, G., Poli, A., Spinsante, S., & Cecchi, S. (2023). Skin conductance under acoustic stimulation: Analysis by a portable device. In *Lecture Notes of the Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering, LNICST*, 456, 62–78. https://doi.org/10.1007/978-3-031-28663-6_6

- [25] Carlson, N. R. (2021). *Physiology of Behaviour* (13th ed.). Pearson Higher Ed.
- [26] Sequeira, H., Hot, P., Silvert, L., & Delplanque, S. (2009). Electrical autonomic correlates of emotion. *International Journal of Psychophysiology*, 71(1), 50–56. <https://doi.org/10.1016/j.ijpsycho.2008.07.009>
- [27] Uno, T., & Grings, W. W. (1965). Autonomic components of orienting behaviour. *Psychophysiology*, 1(4), 311–321. <https://doi.org/10.1111/j.1469-8986.1965.tb03263.x>
- [28] Critchley, H. D., Tang, J., Glaser, D. butterworth, B., & Dolan, R. J. (2005). Anterior cingulate activity during error and autonomic response. *NeuroImage*, 27(4), 885–895. <https://doi.org/10.1016/j.neuroimage.2005.05.047>
- [29] Braem, S., Coenen, E., Bombeke, K., van Bochove, M. E., & Notebaert, W. (2015). Open your eyes for prediction errors. *Cognitive, Affective, & Behavioural Neuroscience*, 15(2), 374–380. <https://doi.org/10.3758/s13415-014-0333-4>
- [30] Kloosterman, N. A., Meindertsma, T., van Loon, A. M., Lamme, V. A. F., Bonne, Y. S., & Donner, T. H. (2015). Pupil size tracks perceptual content and surprise. *European Journal of Neuroscience*, 41(8), 1068–1078. <https://doi.org/10.1111/ejn.12859>
- [31] Antikainen, J., & Niemi, P. (1983). Neuroticism and the pupillary response to a brief exposure to noise. *Biological Psychology*, 17(2-3), 131–135. [https://doi.org/10.1016/0301-0511\(83\)90013-3](https://doi.org/10.1016/0301-0511(83)90013-3)
- [32] Reason, J. (1990). *Human Error*. Cambridge University Press.
- [33] National Highway Traffic Safety Administration (NHTSA). (2015). *Traffic Safety Facts Annual Report*.
- [34] Wickens, C. D., & Hollands, J. G. (2000). *Engineering Psychology and Human Performance*. Prentice Hall.
- [35] Horrey, W. J., Wickens, C. D., & Consalus, K. P. (2006). Modelling driver's visual attention allocation while interacting with in-vehicle technologies. *Journal of Experimental Psychology: Applied*, 12(2), 67.
- [36] Hankins, T. C., & Wilson, G. F. (1998). A comparison of heart rate, eye activity, EEG, and subjective measures of pilot mental workload during flight. *Aviation, Space, and Environmental Medicine*, 69(4), 360-367.
- [37] Fairclough, S. H. (2009). Fundamentals of physiological computing. *Interacting with Computers*, 21(1-2), 133-145.
- [38] Critchley, H. D. (2002). Electrodermal responses: What happens in the brain. *The Neuroscientist*, 8(2), 132-142. <https://doi.org/10.1177/107385840200800209>
- [39] Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, 124(3), 372-422.

- [40] Lee, H., Jiang, N., & Samuel, S. (2025). Detection of error in static and dynamic visual stimulation via electroencephalogram and eye-tracking systems. *Engineering Applications of Artificial Intelligence*.
- [41] Ojsteršek, T. C., & Topolšek, D. (2019). Eye tracking use in researching driver distraction: A scientometric and qualitative literature review approach. *Journal of Eye Movement Research*, 12(3), 1–30.
- [42] Rai, Y., & Le Callet, P. (2018). Visual attention, visual salience, and perceived interest in multimedia applications. In *Academic Press Library in Signal Processing: Image and Video Processing and Analysis and Computer Vision* (Vol. 6, pp. 113–161).
- [43] Gannouni, S., Aledaily, A., Belwafi, K., & Aboalsamh, H. (2021). Emotion detection using electroencephalography signals and a zero-time windowing-based epoch estimation and relevant electrode identification. *Scientific Reports*, 11(1), 1–17. <https://doi.org/10.1038/s41598-020-74509-y>
- [44] Hajcak, G., McDonald, N., & Simons, R. F. (2003). To err is autonomic: Error-related brain potentials, ANS activity, and post-error compensatory behaviour. *Psychophysiology*, 40(6), 895–903.
- [45] Das Chakladar, D., & Roy, P. P. (2024). Cognitive workload estimation using physiological measures: a review. *Cognitive Neurodynamics*, 18(4), 1445-1465.
- [46] Longo, L., Wickens, C. D., Hancock, G., & Hancock, P. A. (2022). Human mental workload: A survey and a novel inclusive definition. *Frontiers in psychology*, 13, 883321.
- [47] Galy, E., Cariou, M., & Mélan, C. (2012). What is the relationship between mental workload factors and cognitive load types?. *International journal of psychophysiology*, 83(3), 269-275.
- [48] Van Acker, B. B., Parmentier, D. D., Vlerick, P., & Saldien, J. (2018). Understanding mental workload: from a clarifying concept analysis toward an implementable framework. *Cognition, technology & work*, 20, 351-365.
- [49] Zhang, Y., Owechko, Y., & Zhang, J. (2004, October). Driver cognitive workload estimation: A data-driven perspective. In *Proceedings. The 7th International IEEE Conference on Intelligent Transportation Systems (IEEE Cat. No. 04TH8749)* (pp. 642-647). IEEE.
- [50] Huang, J., Zhang, Q., Zhang, T., Wang, T., & Tao, D. (2024). Assessment of Drivers' Mental Workload by Multimodal Measures during Auditory-Based Dual-Task Driving Scenarios. *Sensors*, 24(3), 1041.
- [51] Kunasegaran, K., Ismail, A. M. H., Ramasamy, S., Gnanou, J. V., Caszo, B. A., & Chen, P. L. (2023). Understanding mental fatigue and its detection: a comparative analysis of assessments and tools. *PeerJ*, 11, e15744.

- [52] Langner, R., Steinborn, M. B., Chatterjee, A., Sturm, W., & Willmes, K. (2010). Mental fatigue and temporal preparation in simple reaction-time performance. *Acta psychologica*, 133(1), 64-72.
- [53] Hart, S. G. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Human Mental Workload*. Elsevier.
- [54] Liu, Y., Wang, X., & Li, Z. (2016). Time–frequency distribution properties of event-related potentials in mental fatigue induced by visual memory tasks. *Biomedical Signal Processing and Control*, 29, 99–107.
- [55] Królak, A., & Strumiłło, P. (2007). Comparison of eyeblink monitoring and EEG signal analysis for mental fatigue assessment. *Medical Engineering & Physics*, 29(6), 683-691.
- [56] Smith, S. S., et al. (2019). Comparing the effects of three cognitive tasks on indicators of mental fatigue. *Psychophysiology*, 56(6), e13326.
- [57] Xia, Y., et al. (2012). Mental fatigue assessment based on physiological signals. *Computers in Biology and Medicine*, 42(3), 209–217.
- [58] Zhang, W., et al. (2011). Study on visual display terminal mental fatigue status based on physiological information. *Chinese Journal of Medical Instrumentation*, 35(5), 328-331.
- [59] Fan, J., & Smith, A. P. (2017). The impact of workload and fatigue on performance. In *Human Mental Workload: Models and Applications: First International Symposium, H-WORKLOAD 2017, Dublin, Ireland, June 28-30, 2017, Revised Selected Papers 1* (pp. 90-105). Springer International Publishing.
- [60] Xiong, Q., Luo, F., Chen, Y., Duan, Y., Huang, J., Liu, H., ... & Li, R. (2024). Factors influencing fatigue, mental workload and burnout among Chinese health care workers during public emergencies: an online cross-sectional study. *BMC nursing*, 23(1), 428.
- [61] Bakhshi, E., Mazloumi, A., & Hoseini, S. M. (2019). Relationship between mental fatigue and mental workload among nurses. *Zahedan Journal of Research in Medical Sciences*, 21(1).
- [62] Mahdavi, N., Tapak, L., Darvishi, E., Doosti-Irani, A., & Shafiee Motlagh, M. (2024). Unraveling the interplay between mental workload, occupational fatigue, physiological responses and cognitive performance in office workers. *Scientific Reports*, 14(1), 17866.
- [63] Chikhi, S., Matton, N., & Blanchet, S. (2022). EEG power spectral measures of cognitive workload: A meta-analysis. *Psychophysiology*, 59(6), e14009.
- [64] Katmah, R., Al-Shargie, F., Tariq, U., Babiloni, F., Al-Mughairbi, F., & Al-Nashash, H. (2021). A review on mental stress assessment methods using EEG signals. *Sensors*, 21(15), 5043.
- [65] Jensen, O., Gelfand, J., Kounios, J., & Lisman, J. E. (2002). Oscillations in the alpha band (9–12 Hz) increase with memory load during retention in a short-term memory task. *Cerebral cortex*, 12(8), 877-882.

- [66] Klimesch, W. (2012). Alpha-band oscillations, attention, and controlled access to stored information. *Trends in cognitive sciences*, 16(12), 606-617.
- [67] Michels, L., Bucher, K., Lüchinger, R., Klaver, P., Martin, E., Jeanmonod, D., & Brandeis, D. (2010). Simultaneous EEG-fMRI during a working memory task: modulations in low and high frequency bands. *PloS one*, 5(4), e10298.
- [68] Palva, S., & Palva, J. M. (2007). New vistas for α -frequency band oscillations. *Trends in neurosciences*, 30(4), 150-158.
- [69] Gevins, A., Smith, M. E., McEvoy, L., & Yu, D. (1997). High-resolution EEG mapping of cortical activation related to working memory: Effects of task difficulty, type of processing, and practice. *Cerebral Cortex (New York, NY: 1991)*, 7(4), 374–385 10.1093/cercor/7.4.374
- [70] Lei, S., & Roetting, M. (2011). Influence of task combination on EEG spectrum modulation for driver workload estimation. *Human Factors*, 53(2), 168–179.
- [71] Chen, Y., & Huang, X. (2016). Modulation of alpha and beta oscillations during an n-back task with varying temporal memory load. *Frontiers in psychology*, 6, 2031.
- [72] Kakkos, I., Dimitrakopoulos, G. N., Sun, Y., Yuan, J., Matsopoulos, G. K., Bezerianos, A., & Sun, Y. (2021). EEG fingerprints of task-independent mental workload discrimination. *IEEE Journal of Biomedical and Health Informatics*, 25(10), 3824-3833.
- [73] Puma, S., Matton, N., Paubel, P. V., Raufaste, É., & El-Yagoubi, R. (2018). Using theta and alpha band power to assess cognitive workload in multitasking environments. *International Journal of Psychophysiology*, 123, 111-120.
- [74] Arnau, S., Brümmer, T., Liegel, N., & Wascher, E. (2021). Inverse effects of time-on-task in task-related and task-unrelated theta activity. *Psychophysiology*, 58(6), e13805.
- [75] Puma, S., Matton, N., Paubel, P.-V., Raufaste, É., & El-Yagoubi, R. (2018). Using theta and alpha band power to assess cognitive workload in multitasking environments. *International Journal of Psychophysiology*, 123, 111–120.
- [76] Kornblith, S., Buschman, T. J., & Miller, E. K. (2016). Stimulus load and oscillatory activity in higher cortex. *Cerebral Cortex*, 26(9), 3772-3784.
- [77] Proskovec, A. L., Heinrichs-Graham, E., & Wilson, T. W. (2019). Load modulates the alpha and beta oscillatory dynamics serving verbal working memory. *NeuroImage*, 184, 256-265.
- [78] Alberdi, A., Aztiria, A., & Basarab, A. (2016). Towards an automatic early stress recognition system for office environments based on multimodal measurements: A review. *Journal of Biomedical Informatics*, 53, 34-45.
- [79] Andersson, S. (2019). *Universal Design Of Smartphone User interface Interaction in Stressful Situations* (Master's thesis, OsloMet-Oslo Metropolitan University).

- [80] Kakaria, S., & Bigné, E. (2023). CONSUMER BEHAVIOUR IN ELECTRONIC AND VIRTUAL COMMERCE: MERGING BEHAVIOURAL AND NEUROPHYSIOLOGICAL PERSPECTIVES.
- [81] Halbig, A., & Latoschik, M. E. (2021). A systematic review of physiological measurements, factors, methods, and applications in virtual reality. *Frontiers in Virtual Reality*, 2, 694567.
- [82] Sukumaran, A., & Manoharan, A. (2023). A survey on automatic engagement recognition methods: online and traditional classroom. *Indonesian Journal of Electrical Engineering and Computer Science*, 30(2), 1178-1191.
- [83] Giannakakis, G., Pediaditis, M., Manousos, D., Kazantzaki, E., Chiarugi, F., Simos, P. G., ... & Tsiknakis, M. (2017). Stress and anxiety detection using facial cues from videos. *Biomedical Signal Processing and Control*, 31, 89-101.
- [84] Utami, P., Hartanto, R., & Soesanti, I. (2022). A Brief Study of The Use of Pattern Recognition in Online Learning: Recommendation for Assessing Teaching Skills Automatically Online Based. *ELINVO (Electronics, Informatics, and Vocational Education)*, 7(1), 48-62.
- [85] McDuff, D., El Kaliouby, R., Cohn, J. F., & Picard, R. W. (2014). Predicting ad liking and purchase intent: Large-scale analysis of facial responses to ads. *IEEE Transactions on Affective Computing*, 6(3), 223-235.
- [86] Borghini, G. astolfi, L., Vecchiato, G., Mattia, D., & Babiloni, F. (2014). Measuring neurophysiological signals in aircraft pilots and car drivers for the assessment of mental workload, fatigue, and drowsiness. *Neuroscience & Biobehavioural Reviews*, 44, 58-75.
- [87] Fallahi, M., Motamedzade, M., Heidarimoghadam, R., Soltanian, A. R., & Miyake, S. (2016). Effects of mental workload on physiological and subjective responses during traffic density monitoring: A field study. *Applied ergonomics*, 52, 95-103.
- [88] Jap, B. T., Lal, S., Fischer, P., & Bekiaris, E. (2009). Using EEG spectral components to assess algorithms for detecting fatigue. *Expert Systems with Applications*, 36(2), 2352-2359.
- [89] Makeig, S., Debener, S., Onton, J., & Delorme, A. (2004). Mining event-related brain dynamics. *Trends in cognitive sciences*, 8(5), 204-210.
- [90] Shaffer, F., & Ginsberg, J. P. (2017). An overview of heart rate variability metrics and norms. *Frontiers in Public Health*, 5, 258. <https://doi.org/10.3389/fpubh.2017.00258>
- [91] Critchley, H. D. (2002). Electrodermal responses: What happens in the brain. *The Neuroscientist*, 8(2), 132-142. <https://doi.org/10.1177/107385840200800209>
- [92] Marshall, S. P. (2002, September). The index of cognitive activity: Measuring cognitive workload. In *Proceedings of the IEEE 7th conference on Human Factors and Power Plants* (pp. 7-7). IEEE.
- [93] Abraham, H., Lee, C., Brady, S., Fitzgerald, C., Mehler, B., Reimer, B., & Coughlin, J. F. (2017, January). Autonomous vehicles and alternatives to driving: trust, preferences, and effects of age.

- In *Proceedings of the transportation research board 96th annual meeting* (pp. 8-12). Washington, DC: Transportation Research Board.
- [94] Borowsky, A., Shinar, D., & Oron-Gilad, T. (2010). Age, skill, and hazard perception in driving. *Accident analysis & prevention*, 42(4), 1240-1249.
- [95] Choi, J. K., & Ji, Y. G. (2015). Investigating the importance of trust on adopting an autonomous vehicle. *International Journal of Human-Computer Interaction*, 31(10), 692-702.
- [96] He, X., Stapel, J., Wang, M., & Happee, R. (2022). Modelling perceived risk and trust in driving automation reacting to merging and braking vehicles. *Transportation research part F: traffic psychology and behaviour*, 86, 178-195.
- [97] Molnar, L. J., Ryan, L. H., Pradhan, A. K., Eby, D. W., Louis, R. M. S., & Zakrajsek, J. S. (2018). Understanding trust and acceptance of automated vehicles: An exploratory simulator study of transfer of control between automated and manual driving. *Transportation research part F: traffic psychology and behaviour*, 58, 319-328.
- [98] Park, S. Y., Moore, D. J., & Sirkin, D. (2020, April). What a driver wants: User preferences in semi-autonomous vehicle decision-making. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (pp. 1-13).
- [99] Endsley, M. R. (2006). Expertise and situation awareness. *The Cambridge handbook of expertise and expert performance*, 633-651.
- [100] Endsley, M. R. (2017). Direct measurement of situation awareness: Validity and use of SAGAT. In *Situational awareness* (pp. 129-156). Routledge.
- [101] Nordhoff, S., Van Arem, B., Merat, N., Madigan, R., Ruhrort, L., Knie, A., & Happee, R. (2017, June). User acceptance of driverless shuttles running in an open and mixed traffic environment. In *12th ITS European Congress* (pp. 19-22). Strasbourg, France.
- [102] Weiss, E., Fisher Thiel, M., Sultana, N., Hannan, C., & Seacrist, T. (2018). Advanced driver assistance systems for teen drivers: Teen and parent impressions, perceived need, and intervention preferences. *Traffic injury prevention*, 19(sup1), S120-S124.
- [103] Camara, F., Dickinson, P., & Fox, C. (2021). Evaluating pedestrian interaction preferences with a game theoretic autonomous vehicle in virtual reality. *Transportation research part F: traffic psychology and behaviour*, 78, 410-423.
- [104] Forster, Y., Hergeth, S., Naujoks, F., Beggiato, M., Krems, J. F., & Keinath, A. (2019). Learning to use automation: Behavioural changes in interaction with automated driving systems. *Transportation research part F: traffic psychology and behaviour*, 62, 599-614.

- [105] Bellem, H., Thiel, B., Schrauf, M., & Krems, J. F. (2018). Comfort in automated driving: An analysis of preferences for different automated driving styles and their dependence on personality traits. *Transportation research part F: traffic psychology and behaviour*, 55, 90-100.
- [106] Bahrami, S., & Roorda, M. (2022). Autonomous vehicle parking policies: A case study of the City of Toronto. *Transportation Research Part A: Policy and Practice*, 155, 283-296.
- [107] Templeton, B. (2019). How self-driving cars can figure out parking. *Forbes*. <https://www.forbes.com/sites/bradtempleton/2019/06/25/how-self-driving-cars-can-figure-out-parking/?sh=3026765512a2>
- [108] Bosch. (2022). Bosch automated valet parking. *Bosch Mobility Solutions*. <https://www.bosch-mobility-solutions.com/en/solutions/parking/automated-valet-parking/>
- [109] Manawadu, U. E., Ishikawa, M., Kamezaki, M., & Sugano, S. (2015). Analysis of preference for autonomous driving under different traffic conditions using a driving simulator. *Journal of Robotics and Mechatronics*, 27(6), 660-670.
- [110] Ziefle, M., & Bay, S. (2006). How to overcome disorientation in mobile phone menus: A comparison of two different types of navigation aids. *Human-Computer Interaction*, 21(4), 393-433.
- [111] Leiva, L. A., Ttraver, J., Kawala-Sterniuk, A., & Ruotsalo, T. (2024). Modelling User Preferences via Brain-Computer Interfacing. *arXiv preprint arXiv:2405.09691*.
- [112] Kopecka, H. such, J., & Luck, M. (2024). Preferences for AI Explanations Based on Cognitive Style and Socio-Cultural Factors. *Proceedings of the ACM on Human-Computer Interaction*, 8(CSCW1), 1-32.
- [113] Miceli, S., de Palo, V., Monacis, L., Cardaci, M., & Sinatra, M. (2018). The Italian version of the Cognitive Style Indicator and its association with decision-making preferences. *Creativity Research Journal*, 30(1), 85-94.
- [114] Bunt, A., Conati, C., & McGrenere, J. (2007, January). Supporting interface customization using a mixed-initiative approach. In *Proceedings of the 12th international conference on Intelligent user interfaces* (pp. 92-101).
- [115] Dieciuc, M., Zhang, S., Gray, N., Lustria, M. L. A., Carr, D. C., Charness, N., & Boot, W. R. (2024). Terms of engagement: Understanding the motivations, preferences, and attitudes of older adults toward mobile cognitive assessment and training. *The Gerontologist*, 64(2), gnad048.
- [116] Al-Samarraie, H., Selim, H., Teo, T., & Zaqout, F. (2017). Isolation and distinctiveness in the design of e-learning systems influence user preferences. *Interactive Learning Environments*, 25(4), 452-466.

- [117] Ge, Y., He, J., Pang, H., Hu, Y., & Gong, Y. (2019). The Implicit Association Test on Place-of-Origin Stereotyping for Mobile Phone Brands Based on User Preferences Data Processing. *American Journal of Industrial and Business Management*, 9(8), 1699-1711.
- [118] Cassarino, M., Bantry-White, E., & Setti, A. (2019). Cognitive and sensory dimensions of older people's preferences of outdoor spaces for walking: A survey study in Ireland. *International journal of environmental research and public health*, 16(8), 1340.
- [119] Moses-Payne, M. E., Lee, D. G., & Roiser, J. P. (2024). Do adolescents use choice to learn about their preferences? Development of value refinement and its associations with depressive symptoms in adolescence. *Child Development*.
- [120] Rodríguez, D. Z., Wang, Z., Rosa, R. L., & Bressan, G. (2014). The impact of video-quality-level switching on user quality of experience in dynamic adaptive streaming over HTTP. *EURASIP Journal on Wireless Communications and Networking*, 2014, 1-15.
- [121] Khushaba, R. N., Wise, C., Kodagoda, S., & Louviere, J. (2013). Integrating Eye-Tracking and Wireless Electroencephalogram (EEG) in Consumer Neuroscience. In *International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE.
- [122] Bezugam, S. S., Majumdar, S., Ralekar, C., & Gandhi, T. K. (2021). Efficient video summarization framework using eeg and eye-tracking signals. *arXiv preprint arXiv:2101.11249*.
- [123] Zhu, L., & Lv, J. (2023). Review of studies on user research based on EEG and eye tracking. *Applied Sciences*, 13(11), 6502.
- [124] Yang, J., Liang, N., Pitts, B. J., Prakah-Asante, K. O., Curry, R., Blommer, M., ... & Yu, D. (2023). Multimodal sensing and computational intelligence for situation awareness classification in autonomous driving. *IEEE Transactions on Human-Machine Systems*, 53(2), 270-281.
- [125] Norouzi, N., Kim, K., Hochreiter, J., Lee, M., Daher, S., Bruder, G., & Welch, G. (2018, November). A systematic survey of 15 years of user studies published in the intelligent virtual agents conference. In *Proceedings of the 18th international conference on intelligent virtual agents* (pp. 17-22).
- [126] Wang, I., Smith, J., & Ruiz, J. (2019, May). Exploring virtual agents for augmented reality. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (pp. 1-12).
- [127] Shriram, R., Sundhararajan, M., & Daimiwal, N. (2013). EEG-based cognitive workload assessment for maximum efficiency. *International Organization of Scientific Research Journal*, 7, 34-38.
- [128] Thomopoulos, G. A., Lyras, D. P., & Fidas, C. A. (2024). A systematic review and research challenges on phishing cyberattacks from an electroencephalography and gaze-based perspective. *Personal and Ubiquitous Computing*, 1-22.

- [129] Willis, J., & Todorov, A. (2006). First impressions: Making up your mind after a 100-ms exposure to a face. *Psychological science*, 17(7), 592-598.
- [130] Bar, M., Neta, M., & Linz, H. (2006). Very first impressions. *Emotion*, 6(2), 269.
- [131] Green, M. (2000). "How long does it take to stop?" Methodological analysis of driver perception-brake times. *Transportation human factors*, 2(3), 195-216.
- [132] Caird, J. K., Chisholm, S. L., & Lockhart, J. (2008). Do in-vehicle advanced signs enhance older and younger drivers' intersection performance? Driving simulation and eye movement results. *International journal of human-computer studies*, 66(3), 132-144.
- [133] Lerner, N. D. (1993, October). Brake perception-reaction times of older and younger drivers. In *Proceedings of the human factors and ergonomics society annual meeting* (Vol. 37, No. 2, pp. 206-210). Sage CA: Los Angeles, CA: SAGE Publications.
- [134] Teodorovicz, T., Kun, A. L., Sadun, R., & Shaer, O. (2022). Multitasking while driving: A time use study of commuting knowledge workers to assess current and future uses. *International Journal of Human-Computer Studies*, 162, 102789.
- [135] Horrey, W. J., & Wickens, C. D. (2006). Examining the impact of cell phone conversations on driving using meta-analytic techniques. *Human factors*, 48(1), 196-205.
- [136] McGehee, D. V., Mazzae, E. N., & Baldwin, G. S. (2000, July). Driver reaction time in crash avoidance research: Validation of a driving simulator study on a test track. In *Proceedings of the human factors and ergonomics society annual meeting* (Vol. 44, No. 20, pp. 3-320). Sage CA: Los Angeles, CA: Sage Publications
- [137] Critchley, H. D. (2002). Electrodermal responses: what happens in the brain. *The Neuroscientist*, 8(2), 132-142.
- [138] Gonzalez, Cleotilde, et al. "A cognitive modelling account of simultaneous learning and fatigue effects." *Cognitive Systems Research* 12.1 (2011): 19-32.
- [139] Brumby, Duncan P., et al. "Computational models of user multitasking." *Computational interaction design* (2018): 341-362.
- [140] Neuroelectrics, Enobio® technical specifications (2022)
- [141] Ergoneers Dikablis Glasses 3 Eye Tracker, <https://www.jalimedical.com/ergoneers-dikablis-glasses-eye-tracker>
- [142] iMotions. (n.d.). *Empatica E4 EDA/GSR Wristband*. iMotions. Retrieved April 10, 2023, from <https://imotions.com/products/hardware/empatica-e4-eda-gsr-wristband>
- [143] Kakkanatt, S. (2020). Lazy Predict: A tool for quick comparison of machine learning models. *GitHub*. <https://github.com/shankarpandala/lazypredict>

- [144] Guttmann-Flury, E., Sheng, X., & Zhu, X. (2023). Channel selection from source localization: A review of four EEG-based brain–computer interfaces paradigms. *Behaviour Research Methods*, 55(4), 1980-2003.
- [145] Spüler, M., & Niethammer, C. (2015). Error-related potentials during continuous feedback: using EEG to detect errors of different type and severity. *Frontiers in human neuroscience*, 9, 155.
- [146] Brázdil, M., Roman, R., Falkenstein, M., Daniel, P., Jurák, P., & Rektor, I. (2002). Error processing—evidence from intracerebral ERP recordings. *Experimental Brain Research*, 146, 460-466.
- [147] Lopes-Dias, C., Sburlea, A. I., & Müller-Putz, G. R. (2019). Online asynchronous decoding of error-related potentials during the continuous control of a robot. *Scientific reports*, 9(1), 17596.
- [148] Fogelson, N., Shah, M., Scabini, D., & Knight, R. T. (2009). Prefrontal cortex is critical for contextual processing: evidence from brain lesions. *Brain*, 132(11), 3002-3010.
- [149] Duncan-Johnson, C. C., & Donchin, E. (1982). The P300 component of the event-related brain potential as an index of information processing. *Biological psychology*, 14(1-2), 1-52.
- [150] Neill, R. (2024). Multitasking, an EEG experiment: comparative analysis of cognitive workload during demanding stimuli presentation. *Neural Computing and Applications*, 1-17.
- [151] Boksem, M. A., & Smidts, A. (2015). Brain responses to movie trailers predict individual preferences for movies and their population-wide commercial success. *Journal of Marketing Research*, 52(4), 482-492.
- [152] Agarwal, M., & Sivakumar, R. (2019, June). Cerebro: A wearable solution to detect and track user preferences using brainwaves. In *The 5th ACM Workshop on Wearable Systems and Applications* (pp. 47-52).
- [153] Liu, R., Qi, S., Hao, S., Lian, G., & Luo, Y. (2023). Using electroencephalography to analyse drivers' different cognitive workload characteristics based on on-road experiment. *Frontiers in psychology*, 14, 1107176.
- [154] Klimesch, W. (1999). EEG alpha and theta oscillations reflect cognitive and memory performance: A review and analysis. *Brain Research Reviews*, 29(2-3), 169-195.
- [155] Pomplun, M., Sunkara, S., Fairley, A. V., & Xiao, M. (2023). Using pupil size as a measure of cognitive workload in video-based eye-tracking studies. *Journal of Vision*, 23(7), 9-18.
- [156] Khedher, A. B., Jraidi, I., & Frasson, C. (2019, May). Predicting learners' performance using EEG and eye tracking features. In *The thirty-second international flairs conference*.
- [157] Abdurrahman, U. A., Zheng, L., & Yeh, S. (2022). Cognitive workload evaluation of landmarks and routes using virtual reality. *Journal of Cognitive Neuroscience*, 34(1), 35-46.

- [158] Li, X., Rakotonirainy, A., Yan, X., & Zhang, Y. (2018). Driver's visual performance in rear-end collision avoidance process under the influence of cell phone use. *Accident Analysis & Prevention*, 115, 126-137.
- [159] Kim, H. G., Cheon, E. J., Bai, D. S., Lee, Y. H., & Koo, B. H. (2018). Stress and heart rate variability: A meta-analysis and review of the literature. *Psychiatry Investigation*, 15(3), 235-245.
- [160] Vrijkotte, T. G. M., van Doornen, L. J. P., & de Geus, E. J. C. (2000). Effects of work stress on ambulatory blood pressure, heart rate, and heart rate variability. *Hypertension*, 35(4), 880-886.
- [161] Luque-Casado, A., Perales, J. C., Cárdenas, D., & Sanabria, D. (2016). Heart rate variability and cognitive processing: The autonomic response to task demands. *Biological Psychology*, 113, 83-90.
- [162] Jia, Y., Wang, Q., Shi, X., & Li, H. (2018). Perceived user preferences and usability evaluation of mainstream wearable devices for health monitoring. *PeerJ*, 6, e5350.
- [163] Boucsein, W. (2012). *Electrodermal Activity* (2nd ed.). Springer Science & Business Media.
- [164] Nasoz, F., Alvarez, K., Lisetti, C. L., & Finkelstein, N. (2004). Emotion recognition from physiological signals using wireless sensors for presence technologies. *Cognition, Technology & Work*, 6(1), 4-14.
- [165] Fairclough, S. H., & Venables, L. (2006). Prediction of subjective states from psychophysiology: A multivariate approach. *Biological Psychology*, 71(1), 100-110.
- [166] Calvin, K. L., & Duffy, V. G. (2007). Development of a facial skin temperature-based methodology for non-intrusive mental workload measurement. *Occupational Ergonomics*, 7, 83-94.
- [167] Pavlidis, I., Tsiamyrtzis, P., Shastri, D., Wesley, A., Zhou, Y., Lindner, P., & Bas, B. (2012). Fast by nature—How stress patterns define human experience and performance in dexterous tasks. *Scientific Reports*, 2, 1-9.
- [168] Yeom, D. J., & Delogu, F. (2019). Thermal preferences and cognitive performance estimation via user's physiological responses. *ARCC Conference Repository*.
- [169] Wu, Y., & Cao, B. (2022). Recognition and prediction of individual thermal comfort requirement based on local skin temperature. *Journal of Building Engineering*, 49, 104025.
- [170] Moctezuma, L. A., & Molinas, M. (2020). Classification of low-density EEG for epileptic seizures by energy and fractal features based on EMD. *Journal of Biomedical Research*, 34(3), 180-190.
- [171] Amin, H. U., Yusoff, M. Z., & Abdul Majeed, A. P. P. (2015). Feature extraction and classification for EEG signals using wavelet transform and machine learning techniques. *Australasian Physical & Engineering Sciences in Medicine*, 38, 139-149.
- [172] Xiang, F., Ding, L., Feng, Z., & Zhang, X. (2024). Using pupil diameter for psychological resilience assessment in medical students based on SVM and SHAP model. *IEEE Journal of Biomedical and Health Informatics*.

- [173] Strauch, C., Weichwald, S., & Walther, S. (2021). Decoding binary decisions under differential target probabilities from pupil dilation: A random forest approach. *Journal of Vision*, 21(7), 6-6.
- [174] Naik, R., Kroll, N., & Schemmel, S. (2022). The measurement of cognitive workload in surgery using pupil metrics: A systematic review and narrative analysis. *Journal of Surgical Research*, 280, 258-272.
- [175] Sharma, H., Kulkarni, R., & Wagh, S. (2021). Machine learning-based analysis of operator pupillary response to assess cognitive workload in clinical ultrasound imaging. *Computers in Biology and Medicine*, 135, 104589.
- [176] Stewart, C. (2024). *Understanding Disease Through Remote Monitoring Technology* (Doctoral dissertation). King's College London.
- [177] Alberdi, A., Aztiria, A., & Basarab, A. (2016). Towards an automatic early stress recognition system for office environments based on multimodal measurements: A review. *Journal of Biomedical Informatics*, 59, 49-75.
- [178] Plöchl, M., Ossandón, J. P., & König, P. (2012). Combining EEG and eye tracking: Identification, characterization, and correction of eye movement artifacts in electroencephalographic data. *Frontiers in Human Neuroscience*, 6, 278.
- [179] Croft, R. J., & Barry, R. J. (2000). Removal of ocular artifact from the EEG: A review. *Neurophysiologie Clinique/Clinical Neurophysiology*, 30(1), 5-19.
- [180] Joyce, C. A., Gorodnitsky, I. F., & Kutas, M. (2004). Automatic removal of eye movement and blink artifacts from EEG data using blind component separation. *Psychophysiology*, 41(2), 313-325.
- [181] Jung, T.-P., Makeig, S., Humphries, C., Lee, T.-W., McKeown, M. J., Iragui, V., & Sejnowski, T. J. (2000). Removing electroencephalographic artifacts by blind source separation. *Psychophysiology*, 37(2), 163-176.
- [182] Delorme, A., & Makeig, S. (2004). EEGLAB: An open-source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of Neuroscience Methods*, 134(1), 9-21.
- [183] Mitra, P. P., & Pesaran, B. (1999). Analysis of dynamic brain imaging data. *Biophysical Journal*, 76(2), 691-708.
- [184] The Carpentries Incubator. (2018). *Processing data with EEGLAB*. SDC-EEGLAB. <https://carpentries-incubator.github.io/SDC-EEGLAB/index.html>
- [185] Dimigen, O. (2020). Optimizing the ICA-based removal of ocular EEG artifacts from free viewing experiments. *NeuroImage*, 207, 116117.
- [186] Hampel, F. R. (1974). The influence curve and its role in robust estimation. *Journal of the American Statistical Association*, 69(346), 383-393.

- [187] Salvucci, D. D., & Goldberg, J. H. (2000). Identifying fixations and saccades in eye-tracking protocols. In *Proceedings of the 2000 Symposium on Eye Tracking Research & Applications* (pp. 71-78).
- [188] Salvucci, D. D., & Goldberg, J. H. (2000, November). Identifying fixations and saccades in eye-tracking protocols. In *Proceedings of the 2000 symposium on Eye tracking research & applications* (pp. 71-78).
- [189] Ramos, E. D. (2019). Automatic Prediction of Health Literacy through an eye tracker
- [190] Dupont, L., Antrop, M., & Van Eetvelde, V. (2014). Eye-tracking analysis in landscape perception research: Influence of photograph properties and landscape characteristics. *Landscape research*, 39(4), 417-432.
- [191] Komogortsev, O. V., Gobert, D. V., Jayarathna, S., & Gowda, S. M. (2010). Standardization of automated analyses of oculomotor fixation and saccadic behaviours. *IEEE Transactions on biomedical engineering*, 57(11), 2635-2645.
- [192] Empatica Support. (2024). Empatica E4 wristband product specifications. *Empatica Support*. <https://support.empatica.com/hc/en-us/articles/202581999-E4-wristband-technical-specifications>
- [193] Empatica Support. (2024). Utilizing the PPG/BVP signal. *Empatica Support*. <https://support.empatica.com/hc/en-us/articles/204954639-Utilizing-the-PPG-BVP-signal>
- [194] Draghici, A. E., & Taylor, J. A. (2016). The physiological basis and measurement of heart rate variability in humans. *Journal of Physiological Anthropology*, 35, 1-8.
- [195] Empatica Support. (2024). E4 data - IBI expected signal. *Empatica Support*. <https://support.empatica.com/hc/en-us/articles/360030058011-E4-data-IBI-expected-signal>
- [196] Empatica Support. (2024). What should I know to use EDA data in my experiment? *Empatica Support*. <https://support.empatica.com/hc/en-us/articles/203621955-What-should-I-know-to-use-EDA-data-in-my-experiment>
- [197] Liao, J. C., Shih, W. Y., Huang, K. J., & Fang, W. C. (2013, April). An online recursive ICA based real-time multichannel EEG system on chip design with automatic eye blink artifact rejection. In *2013 International Symposium on VLSI Design, Automation, and Test (VLSI-DAT)* (pp. 1-4). IEEE.
- [198] Pion-Tonachini, L., Hsu, S. H., Makeig, S., Jung, T. P., & Cauwenberghs, G. (2015, August). Real-time eeg source-mapping toolbox (rest): Online ica and source localization. In *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (pp. 4114-4117). IEEE.

- [199] Jafarifarmand, A., Badamchizadeh, M. A., Khanmohammadi, S., Nazari, M. A., & Tazehkand, B. M. (2017). Real-time ocular artifacts removal of EEG data using a hybrid ICA-ANC approach. *Biomedical signal Processing and control*, 31, 199-210.
- [200] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ... & Liu, T. Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30.
- [201] American Association of State Highway and Transportation Officials (AASHTO). (2018). *A Policy on Geometric Design of Highways and Streets* (7th ed.). Washington, D.C.: AASHTO. Available at: Best Materials.
- [202] Guo, F., Li, M., Chen, Y., Xiong, J., & Lee, J. (2019). Effects of highway landscapes on drivers' eye movement behaviour and emergency reaction time: a driving simulator study. *Journal of advanced transportation*, 2019(1), 9897831.
- [203] National Highway Traffic Safety Administration (NHTSA). (1995). *Federal Motor Vehicle Safety Standards; Air Brake Systems (FMVSS No. 121)*. Code of Federal Regulations, Title 49, Part 571. Available at: Electronic Code of Federal Regulations (eCFR).
- [204] Neçaibia, A., Ladaci, S., & Mekhilef, S. (2015). ABS braking control enhancement via fractional order extremum seeking method. *Journal of Automation Systems Engineering*, 9(1), 23-36.
- [205] U.S. Department of Transportation, Federal Highway Administration. (2009). *Chapter 4: Engineering and technical measures*. Speed Concepts: Informational Guide
- [206] Taoka, G. T. J. I. (1989). Break reaction times of unalerted drivers. *ITE journal*, 59(3), 19-21.
- [207] Ibanez, A., Melloni, M., Huepe, D., Helgiu, E., Rivera-Rei, A., Canales-Johnson, A., ... & Moya, A. (2012). What event-related potentials (ERPs) bring to social neuroscience?. *Social neuroscience*, 7(6), 632-649.
- [208] Tanaka, H., & Jiang, P. (2024). P1, N170, and N250 Event-related Potential Components Reflect Temporal Perception Processing in Face and Body Personal Identification. *Journal of Cognitive Neuroscience*, 1-17.
- [209] Sieber, M., & Färber, B. (2016, June). Driver perception and reaction in collision avoidance: Implications for ADAS development and testing. In *2016 IEEE Intelligent Vehicles Symposium (IV)* (pp. 239-245). IEEE.
- [210] Wang, W., Xi, J., & Zhao, D. (2018). Learning and inferring a driver's braking action in car-following scenarios. *IEEE Transactions on Vehicular Technology*, 67(5), 3887-3899.
- [211] https://link.springer.com/article/10.1007/s40890-017-0054-9?utm_source=chatgpt.com
- [212] Swedler, D. I., Ali, B., Hoffman, R., Leonardo, J., Romano, E., & Miller, T. R. (2024). Injury and fatality risks for child pedestrians and cyclists on public roads. *Injury epidemiology*, 11(1), 15.

- [213] Peden, M. M. (Ed.). (2008). World report on child injury prevention.
- [214] Stoker, P., Garfinkel-Castro, A., Khayesi, M., Odero, W., Mwangi, M. N., Peden, M., & Ewing, R. (2015). Pedestrian safety and the built environment: a review of the risk factors. *Journal of Planning Literature*, 30(4), 377-392.
- [215] Raford, N., & Ragland, D. (2004). Space syntax: Innovative pedestrian volume modeling tool for pedestrian safety. *Transportation research record*, 1878(1), 66-74.
- [216] Xu, R., Wang, Y., Wang, N., Shi, X., Meng, L., & Ming, D. (2020, July). The effect of static and dynamic visual stimulations on error-evoked brain responses. In *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)* (pp. 2877-2880). IEEE.
- [217] Singh, A., Murzello, Y., Lee, H., Abdalla, S., & Samuel, S. (2024). Moral decision making: Explainable insights into the role of working memory in autonomous driving. *Machine Learning with Applications*, 18, 100599.
- [218] Nakashima, R., & Yokosawa, K. (2018). To see dynamic change: continuous focused attention facilitates change detection, but the effect persists briefly. *Visual Cognition*, 26(1), 37-47.
- [219] https://en.wikipedia.org/wiki/Posner_cueing_task?utm_source=chatgpt.com
- [220] Cohen, J. (2013). *Statistical power analysis for the behavioural sciences*. routledge.
- [221] Luck, S. J. (2014). *An introduction to the event-related potential technique*. MIT press.
- [222] Behzadnia, A., Ghoshuni, M., & Chermahini, S. A. (2017). EEG activities and the sustained attention performance. *Neurophysiology*, 49(3), 226-233.
- [223] Li, W., Zhu, G., Jiang, Y., Miao, C., Zhang, G., & Xu, D. (2024). Cortical response characteristics of passive, active, and resistance movements: a multi-channel fNIRS study. *Frontiers in Human Neuroscience*, 18, 1419140.
- [224] Wang, Y., Guo, Y., Wang, J., Liu, Z., & Li, X. (2021). Pupillary response to moving stimuli of different speeds. *Journal of Eye Movement Research*, 14(1).
- [225] Grootjen, J. W., Thalhammer, P., & Kosch, T. (2024). Your Eyes on Speed: Using Pupil Dilation to Adaptively Select Speed-Reading Parameters in Virtual Reality. *Proceedings of the ACM on Human-Computer Interaction*, 8(MHCI), 1-17.
- [226] Kang, O. E., Huffer, K. E., & Wheatley, T. P. (2014). Pupil dilation dynamics track attention to high-level information. *PloS one*, 9(8), e102463.
- [227] Arrighi, P., Bonfiglio, L., Minichilli, F., Cantore, N., Carboncini, M. C., Piccotti, E., ... & Andre, P. (2016). EEG theta dynamics within frontal and parietal cortices for error processing during reaching movements in a prism adaptation study altering visuo-motor predictive planning. *PLoS One*, 11(3), e0150265.

- [228] Tang, R., Kim, J. H., Parker, R., & Jeong, Y. J. (2018). Indicating severity of vehicle accidents using pupil diameter in a driving simulator environment. In *Digital Human Modeling. Applications in Health, Safety, Ergonomics, and Risk Management: 9th International Conference, DHM 2018, Held as Part of HCI International 2018, Las Vegas, NV, USA, July 15-20, 2018, Proceedings 9* (pp. 647-656). Springer International Publishing.
- [229] Yang, X., & Kim, J. H. (2022, June). Pupil and electromyography (EMG) responses to collision warning in a real driving environment. In *International Conference on Human-Computer Interaction* (pp. 415-423). Cham: Springer International Publishing.
- [230] Joshi, H. C., & Kumar, S. (2024). Artificial Intelligence Failures in Autonomous Vehicles: Causes, Implications, and Prevention. *Computer*, 57(11), 18-30.
- [231] Matos, F., Bernardino, J., Durães, J., & Cunha, J. (2024). A Survey on Sensor Failures in Autonomous Vehicles: Challenges and Solutions. *Sensors*, 24(16), 5108.
- [232] Mirzarazi, F., Danishvar, S., & Mousavi, A. (2024). The Safety Risks of AI-Driven Solutions in Autonomous Road Vehicles. *World Electric Vehicle Journal*, 15(10), 438.
- [233] Rendon-Velez, E., Van Leeuwen, P. M., Happee, R., Horvath, I., van der Vegte, W. F., & de Winter, J. C. (2016). The effects of time pressure on driver performance and physiological activity: A driving simulator study. *Transportation research part F: traffic psychology and behaviour*, 41, 150-169.
- [234] Magaña, V. C., Scherz, W. D., Seepold, R., Madrid, N. M., Pañeda, X. G., & Garcia, R. (2020). The effects of the driver's mental state and passenger compartment conditions on driving performance and driving stress. *Sensors*, 20(18), 5274.
- [235] Darzi, A., Gaweesh, S. M., Ahmed, M. M., & Novak, D. (2018). Identifying the causes of drivers' hazardous states using driver characteristics, vehicle kinematics, and physiological measurements. *Frontiers in neuroscience*, 12, 568.
- [236] Abdel-Aty, M., & Ding, S. (2024). A matched case-control analysis of autonomous vs human-driven vehicle accidents. *Nature Communications*, 15(1), 4931.
- [237] Di Lillo, L., Gode, T., Zhou, X., Atzei, M., Chen, R., & Victor, T. (2024). Comparative safety performance of autonomous-and human drivers: A real-world case study of the Waymo Driver. *Heliyon*, 10(14).
- [238] Durlík, I., Miller, T., Kostecka, E., Zwierzewicz, Z., & Łobodzińska, A. (2024). Cybersecurity in Autonomous Vehicles—Are We Ready for the Challenge?. *Electronics*, 13(13), 2654.
- [239] Blinder, C., & Uhler, K., (2024). How safe are autonomous vehicles? 2025. ConsumerAffairs.

- [240] Bagloee, S. A., Tavana, M., Asadi, M., & Oliver, T. (2016). Autonomous vehicles: challenges, opportunities, and future implications for transportation policies. *Journal of modern transportation*, 24, 284-303.
- [241] Lee, H., Murzello, Y., Nabipour, M., & Samuel, S. (2024). Investigation of driver preference for a user-centred design of decision systems in autonomous vehicles, part I: preferences for binary self-driving modes. *Theoretical Issues in Ergonomics Science*, 1-18.
- [242] Lee, H., & Samuel, S. (2024). Impact of Aging on Driver Preferences for Self-Driving Modes and Behaviors in Two Traffic Complexities. *IEEE Access*, 12, 140144-140159.
- [243] Lee, H., & Samuel, S. (2024). Classification of User Preference for Self-Driving Mode and Behaviors of Autonomous Vehicle. *IEEE Transactions on Intelligent Vehicles*.
- [244] Azevedo-Sa, H., Jayaraman, S. K., Esterwood, C. T., Yang, X. J., Robert Jr, L. P., & Tilbury, D. M. (2021). Real-time estimation of drivers' trust in automated driving systems. *International Journal of Social Robotics*, 13(8), 1911-1927.
- [245] Naiseh, M., Clark, J., Akarsu, T., Hanoch, Y., Brito, M., Wald, M., ... & Shukla, P. (2024). Trust, risk perception, and intention to use autonomous vehicles: an interdisciplinary bibliometric review. *AI & SOCIETY*, 1-21.
- [246] Chen, L. C., Sandmann, P., Thorne, J. D., Herrmann, C. S., & Debener, S. (2015). Association of concurrent fNIRS and EEG signatures in response to auditory and visual stimuli. *Brain topography*, 28, 710-725.
- [247] Studenova, A., Forster, C., Engemann, D. A., Hensch, T., Sanders, C., Mauche, N., ... & Nikulin, V. (2023). Event-related modulation of alpha rhythm explains the auditory P300-evoked response in EEG. *Elife*, 12, RP88367.
- [248] Hoogenboom, N., Schoffelen, J. M., Oostenveld, R., Parkes, L. M., & Fries, P. (2006). Localizing human visual gamma-band activity in frequency, time and space. *Neuroimage*, 29(3), 764-773.
- [249] Mate, K. K. (2022). Effects of Visual, Auditory, and Combined Cues on Human Movement and Brain Regions Involved in Perception Action. *McGill Journal of Medicine*, 20(2).
- [250] Ban, S., Lee, Y.J., Yu, K.J., Chang, J.W., Kim, J.H. and Yeo, W.H., 2023. Persistent human-machine interfaces for robotic arm control via gaze and eye direction tracking. *Advanced Intelligent Systems*, 5(7), p.2200408. DOI: 10.1002/AISY.202200408.
- [251] Ciabattoni, L., et al., 2019. ErrP signals detection for safe navigation of a smart wheelchair. *2019 IEEE 23rd International Symposium on Consumer Technologies (ISCT)*, pp.1-5. DOI: 10.1109/ISCT.2019.8900941.

- [252] De, B., et al., 2023. Design of a brain-actuated 2-loop position control of a robot arm for neuro-motor rehabilitation. *2023 8th International Conference on Computers and Devices for Communication (CODEC)*, pp.1–6. DOI: 10.1109/CODEC56744.2023.10122048.
- [253] Pires, G., Castelo-Branco, M., Guger, C. and Cisotto, G., 2022. Editorial: Error-related potentials: Challenges and applications. *Frontiers in Human Neuroscience*, 16, p.984254. DOI: 10.3389/fnhum.2022.984254.
- [254] Stalljann, S., Wöhle, L., Schäfer, J. and Gebhard, M., 2020. Performance analysis of a head and eye motion-based control interface for assistive robots. *Sensors*, 20(24), p.7162. DOI: 10.3390/s20247162.
- [255] Kalinkov, K. and Markova, V., 2022. Preprocessing of PPG and EDA signals for detection of emotional and cognitive states via physiological signals. *Annual Journal of Technical University of Varna, Bulgaria*, 6(2), pp.49–56.
- [256] Sinha, A., et al., 2016. Physiological sensing-based stress analysis during assessment. *2016 IEEE Frontiers in Education Conference (FIE)*, pp.1–4. IEEE.
- [257] Cano, S., et al., 2024. Wearable solutions using physiological signals for stress monitoring on individuals with Autism Spectrum Disorder (ASD): A systematic literature review. *Sensors*, 24(24), p.8137.
- [258] Bamonte, M.F., Risk, M. and Herrero, V., 2024. Determining the optimal window duration to enhance emotion recognition based on galvanic skin response and photoplethysmography signals. *Electronics*, 13(16), p.3333.
- [259] Ayata, D., Yaslan, Y. and Kamasak, M.E., 2018. Emotion-based music recommendation system using wearable physiological sensors. *IEEE Transactions on Consumer Electronics*, 64(3), pp.196–203.
- [260] Zhang, T., Ali, A.E., Wang, C., Hanjalic, A. and Cesar, P., 2021. CorrNet: Fine-grained emotion recognition for video watching using wearable physiological sensors. *Sensors*, 21(1), p.52.
- [261] Dawson, M.E., Schell, A.M. and Courtney, C.G., 2011. The skin conductance response, anticipation, and decision-making. *Journal of Neuroscience, Psychology, and Economics*, 4(2), pp.111–116.
- [262] Lau-Zhu, A., Lau, M.P.H. and McLoughlin, G., 2019. Mobile EEG in research on neurodevelopmental disorders: Opportunities and challenges. *Developmental Cognitive Neuroscience*, 36, p.100635. DOI: 10.1016/j.dcn.2019.100635.
- [263] Rudnicki, W.R., Kurska, M.B. and Jaworska, M., 2017. Stability of feature selection algorithms: A review. *Frontiers in Artificial Intelligence and Applications*, 300, pp.1–14. DOI: 10.3233/978-1-61499-672-9-177

- [264] Amadori, P. V., & Demiris, Y. (2023). User-Aware Multi-Level Cognitive Workload Estimation from Multimodal Physiological Signals. *IEEE Transactions on Cognitive and Developmental Systems*.
- [265] Ma, X., Monfared, R., Grant, R., & Goh, Y. M. (2024). Determining Cognitive Workload Using Physiological Measurements: Pupillometry and Heart-Rate Variability. *Sensors*, 24(6), 2010.
- [266] Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G* Power 3: A flexible statistical power analysis program for the social, behavioural, and biomedical sciences. *Behaviour research methods*, 39(2), 175-191.
- [267] Krejtz, K., Duchowski, A. T., Niedzielska, A., Biele, C., & Krejtz, I. (2018). Eye tracking cognitive load using pupil diameter and microsaccades with fixed gaze. *PloS one*, 13(9), e0203629.
- [268] Momeni, N., Dell'Agnola, F., Arza, A., & Atienza, D. (2019, July). Real-time cognitive workload monitoring based on machine learning using physiological signals in rescue missions. In *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (pp. 3779-3785). IEEE.
- [269] Liang, Z., Wang, Y., Qian, C., Wang, Y., Zhao, C., Du, H., ... & He, Y. (2024). A Driving Simulator Study to Examine the Impact of Visual Distraction Duration from In-Vehicle Displays: Driving Performance, Detection Response, and Mental Workload. *Electronics*, 13(14), 2718.
- [270] Häusser, J. A., Schulz-Hardt, S., Schultze, T., Tomaschek, A., & Mojzisch, A. (2014). Experimental evidence for the effects of task repetitiveness on mental strain and objective work performance. *Journal of Organizational Behavior*, 35(5), 705-721.
- [271] Charles, R. L., & Nixon, J. (2019). Measuring mental workload using physiological measures: A systematic review. *Applied ergonomics*, 74, 221-232.
- [272] M. Malakoutikhah *et al.*, "Comparison of mental workload with N-Back test: A new design for NASA-task load index questionnaire," *International Archives of Health Sciences*, vol. 8, no. 1, pp. 7–13, 2021.
- [273] H. Devos *et al.*, "Psychometric properties of NASA-TLX and index of cognitive activity as measures of cognitive workload in older adults," *Brain Sciences*, vol. 10, no. 12, p. 994, 2020.
- [274] A. Wang *et al.*, "The association between physiological and eye-tracking metrics and cognitive load in drivers: A meta-analysis," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 104, pp. 474–487, 2024.
- [275] National Safety Council. (2024). Fatigued Driving. *National Safety Council*. <https://www.nsc.org/road/safety-topics/fatigued-driver>

- [276] Transportation Services. (2023). Traffic Signal Vehicle and Pedestrian Volumes (with Traffic Volumes at Intersections for All Modes). *City of Toronto Open Data Portal*. <https://open.toronto.ca/dataset/traffic-signal-vehicle-and-pedestrian-volumes/>
- [277] Park, S. Y., Moore, D. J., & Sirkin, D. (2020). What a driver wants: User preferences in semi-autonomous vehicle decision-making. In *Conference on Human Factors in Computing Systems - Proceedings*.
- [278] Billings, C. E. (1996). *Human-Centred Aviation Automation: Principles and Guidelines*.
- [279] Delmas, M., Camps, V., & Lemerrier, C. (2023). Should my automated car drive as I do? Investigating speed preferences of drivengers in various driving conditions. *PLoS ONE*, 18(2), e0281702. <https://doi.org/10.1371/journal.pone.0281702>
- [280] Navarro, J. (2019). A state of science on highly automated driving. *Theoretical Issues in Ergonomics Science*, 20(3), 366-396. <https://doi.org/10.1080/1463922X.2018.1439544>
- [281] Liljamo, T., Liimatainen, H., & Pöllänen, M. (2018). Attitudes and concerns on automated vehicles. *Transportation Research Part F: Traffic Psychology and Behaviour*, 59, 24-44. <https://doi.org/10.1016/j.trf.2018.08.010>
- [282] Lee, D., & Hess, D. J. (2020). Regulations for on-road testing of connected and automated vehicles: Assessing the potential for global safety harmonization. *Transportation Research Part A: Policy and Practice*, 136, 85-98.
- [283] Vecchiato, G., Toppi, J., Astolfi, L., De Vico Fallani, F., Cincotti, F., Mattia, D., ... & Babiloni, F. (2011). Spectral EEG frontal asymmetries correlate with the experienced pleasantness of TV commercial advertisements. *Medical & biological engineering & computing*, 49, 579-583.
- [284] Slanzi, G., Balazs, J. A., & Velásquez, J. D. (2017). Combining eye tracking, pupil dilation and EEG analysis for predicting web users click intention. *Information Fusion*, 35, 51-57.
- [285] Khushaba, R. N., Kodagoda, S., Dissanayake, G., Greenacre, L., Burke, S., & Louviere, J. (2012, June). A neuroscientific approach to choice modelling: Electroencephalogram (EEG) and user preferences. In *The 2012 international joint conference on neural networks (IJCNN)* (pp. 1-8). IEEE.
- [286] Zhu, S., Qi, J., Hu, J., & Hao, S. (2022). A new approach for product evaluation based on integration of EEG and eye-tracking. *Advanced Engineering Informatics*, 52, 101601.
- [287] Erdfold, E., Faul, F., Buchner, A., & Lang, A. G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behaviour Research Methods*, 41(4), 1149-1160. <https://doi.org/10.3758/BRM.41.4.1149>
- [288] MATLAB. (2010). *Version 7.10.0 (R2010a)*. Natick, Massachusetts. Accessed: Jun. 07, 2022. Available: <https://www.mathworks.com/products/matlab.html>

- [289] IBM Corp. (2019). *IBM SPSS Statistics for Windows, Version 26.0*. IBM Corp. Armonk, NY.
<https://www.ibm.com/support/pages/downloading-ibm-spss-statistics-26>
- [290] Pandala, S. R., & Silva, B. (2022). Lazy Predict Project. *The Python Package Index*.
<https://pypi.org/project/lazypredict/>
- [291] Guolin, K., et al. (2017). LightGBM: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems*, 30.
- [292] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785-794).
- [293] Dorogush, A. V., Ershov, V., & Gulin, A. (2018). CatBoost: gradient boosting with categorical features support. *arXiv preprint arXiv:1810.11363*.
- [294] Zöller, M. A., & Huber, M. F. (2021). Benchmark and Survey of Automated Machine Learning Frameworks. *Journal of Artificial Intelligence Research*, 70, 409-472.
- [295] Wu, J., Chen, X. Y., Zhang, H., Xiong, L. D., Lei, H., & Deng, S. H. (2019). Hyperparameter optimization for machine learning models based on Bayesian optimization. *Journal of Electronic Science and Technology*, 17(1), 26-40.
- [296] Keany, E. (2020). BorutaShap: A wrapper feature selection method which combines the Boruta feature selection algorithm with Shapley values. *Zenodo*. <https://zenodo.org/record/4247618>
- [297] Kursu, M. B., & Rudnicki, W. R. (2010). Feature selection with the Boruta package. *Journal of Statistical Software*, 36(11), 1-13. <https://doi.org/10.18637/jss.v036.i11>
- [298] Shapley, L. (1953). A value for n-person games. In *Contributions to the Theory of Games II* (pp. 307-317). Princeton University Press. <https://doi.org/10.1515/9781400881970-018>
- [299] Choi, J. K., & Ji, Y. G. (2015). Investigating the importance of trust on adopting an autonomous vehicle. *International Journal of Human-Computer Interaction*, 31(10), 692-702.
<https://doi.org/10.1080/10447318.2015.1070549>
- [300] Lajunen, T., & Sullman, M. J. M. (2021). Attitudes toward four levels of self-driving technology among elder drivers. *Frontiers in Psychology*, 12, 682973.
<https://doi.org/10.3389/fpsyg.2021.682973>
- [301] Spiceworks. (2022). What is ADAS (Advanced Driver Assistance Systems)? *Spiceworks*.
<https://www.spiceworks.com/tech/iot/articles/what-is-adas/>

Appendix A. Lazy Predict: 30 ML models

Model	Brief explanation
AdaBoost Classifier	An ensemble method that combines multiple weak learners (typically decision trees) to create a robust classifier.
Bagging Classifier	An ensemble method that builds multiple models on different subsets of the data and averages their predictions.
Bernoulli Naive Bayes	A version of Naive Bayes suited for binary/Boolean features.
Calibrated Classifier CV	A method that provides calibrated probability estimates for base classifiers.
Decision Tree Classifier	A non-parametric model that splits data based on feature values to make predictions.
Dummy Classifier	A simple baseline classifier that makes predictions using simple rules (e.g., always predicting the most frequent class).
Extra Trees Classifier	An ensemble method similar to Random Forests but with more random splits, often providing better performance.
Gradient Boosting Classifier	An ensemble method that builds trees sequentially, with each tree correcting the errors of the previous one.
K-Nearest Neighbors	A non-parametric method that predicts the class of a data point based on the majority class among its k-nearest neighbours.
Label Propagation	A semi-supervised learning algorithm that spreads label information across the data points based on graph structure.
Label Spreading	A variation of Label Propagation with a different regularization approach for spreading labels.
Linear Discriminant Analysis	A linear classifier that finds a linear combination of features to separate classes.
Linear SVC (Support Vector Classifier)	A support vector machine that seeks to find the best hyperplane that separates classes in a linear fashion.
Logistic Regression	A linear model used for binary classification tasks, predicting the probability of class membership.
MLP Classifier (Multi-layer Perceptron)	A neural network model with multiple layers (Multi-layer Perceptron) used for complex patterns and data.
Nearest Centroid	A simple model that classifies data points based on the nearest centroid of each class.
Passive Aggressive Classifier	An online learning algorithm that adjusts its model to correct mistakes while being aggressive about it.
Perceptron	A simple type of artificial neural network used for binary classification tasks.
Quadratic Discriminant Analysis	A version of LDA that allows for a quadratic decision boundary, handling cases where class covariances are different.

Random Forest Classifier	An ensemble of decision trees where each tree is trained on a random subset of the data, reducing overfitting.
Ridge Classifier	A linear classifier that includes a regularization term to prevent overfitting by penalizing large coefficients.
Ridge Classifier CV	Similar to the Ridge Classifier, it includes cross-validation to find the best regularization parameter.
SGD Classifier	(Stochastic Gradient Descent) A linear model optimized using Stochastic Gradient Descent, efficient for large-scale datasets.
SVC (Support Vector Classifier)	A support vector machine that tries to separate data with the widest possible margin useful for both linear and non-linear classification.
Extra Tree Classifier	A version of the decision tree classifier that randomly selects splits for trees, often resulting in better performance and faster computation.
Gaussian Naive Bayes	A version of Naive Bayes that assumes features follow a Gaussian distribution used for continuous data.
Gaussian Process Classifier	A probabilistic model where predictions are based on the Gaussian process, providing a distribution over possible functions fitting the data.
Hist Gradient Boosting Classifier	An ensemble method like Gradient Boosting but optimized with histogram-based learning allows faster and more scalable training.
LightGBM Classifier	A highly efficient gradient boosting framework that uses tree-based learning, designed for fast training and low memory usage.
XGBoost Classifier	An optimized gradient-boosting model that provides high performance and flexibility is widely used in machine learning competitions.

Appendix B. Extracted Features

Power

Power is a fundamental feature that represents the total energy contained within a specific frequency band of a physiological signal. Mathematically, it is expressed as:

$$P = \sum_{f=f_1}^{f_2} |X(f)|^2$$

where $X(f)$ denotes the signal's Fourier transform within the frequency range $[f_1, f_2]$. This feature is crucial for identifying the dominant frequency components that contribute to the physiological response.

Relative power

Relative power provides a normalized measure of power within a specific frequency band relative to the total power across all frequencies. It is given by:

$$P_{\text{rel}} = \frac{P_{\text{band}}}{P_{\text{total}}}$$

This feature helps highlight the contribution of specific frequency bands to the overall signal, making it useful for identifying shifts in physiological states.

Mean Energy

Mean energy reflects the average intensity of the signal over time calculated as:

$$E_{\text{mean}} = \frac{1}{N} \sum_{i=1}^N x_i^2$$

where x_i represents the signal value at the i -th time point. Mean energy is particularly useful for assessing the general activation level of physiological responses.

Log Energy Entropy

Log energy entropy measures the unpredictability or randomness in the distribution of signal energy defined as:

$$H_{\text{log}} = -\sum_{i=1}^N (x_i^2 \cdot \log(x_i^2))$$

This feature is instrumental in detecting transitions in physiological states as higher entropy typically indicates more complex or variable responses.

Shannon Entropy

Shannon entropy is a classic measure of signal complexity representing the amount of uncertainty or information contained in the signal:

$$H_s = -\sum_{i=1}^N p_i \log(p_i)$$

where p_i is the probability of occurrence of the i -th signal value. This feature is widely used to assess the richness and variability of physiological responses.

Mean

The mean value of the signal provides a measure of the physiological state over time computed as:

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

where x_i represents the signal value at the i -th time point. This simple feature helps in establishing a reference point for analyzing deviations in the signal.

Variance

Variance measures the signal's variability around its mean indicating the extent to which the physiological response fluctuates:

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2$$

where x_i represents the signal value at the i -th time point. A higher variance indicates more dynamic changes in the physiological state which may correlate with heightened cognitive or emotional activity.

Skewness

Skewness quantifies the asymmetry of the signal distribution providing insights into the balance of the physiological response:

$$\gamma_1 = \frac{1}{N} \sum_{i=1}^N \left(\frac{x_i - \mu}{\sigma} \right)^3$$

where x_i represents the signal value at the i -th time point. Positive or negative skewness can reveal tendencies in the physiological response that may be significant for interpreting user behaviour.

Kurtosis

Kurtosis measures the "tailedness" of the signal distribution, indicating the presence of outliers:

$$\gamma_2 = \frac{1}{N} \sum_{i=1}^N \left(\frac{x_i - \mu}{\sigma} \right)^4 - 3$$

where x_i represents the signal value at the i -th time point. This feature is essential for detecting unusual physiological events that could impact user experience or safety.

Hjorth Activity

Hjorth activity is equivalent to the signal's variance and reflects its overall intensity:

$$A = \sigma^2$$

This feature is commonly used in time-series analysis to assess the general activation level of the physiological signal.

Hjorth Mobility

Hjorth mobility assesses the frequency content of the signal, indicating how quickly the signal changes over time:

$$M = \sqrt{\frac{\text{Variance of } \dot{x}(t)}{\text{Variance of } x(t)}}$$

where $x(t)$ represents the signal value at the t -th time point and $\dot{x}(t)$ represents the first order of the signal value. This feature is useful for understanding the temporal dynamics of the physiological response.

Hjorth Complexity

Hjorth complexity measures the complexity of the signal's waveform relative to a sine wave:

$$C = \frac{M(\dot{x}(t))}{M(x(t))}$$

This feature is vital for identifying intricate patterns in physiological signals that may correspond to complex cognitive processes.

Normalized 1st Difference

The normalized first difference captures the rate of change in the signal, helping to identify rapid transitions:

$$\Delta_1 = \frac{x_{i+1} - x_i}{\sigma}$$

This feature is particularly useful for detecting sudden shifts in physiological states.

Normalized 2nd Difference

The normalized second difference provides insights into the acceleration of changes in the signal:

$$\Delta_2 = \frac{x_{i+2} - 2x_{i+1} + x_i}{\sigma}$$

This feature is essential for understanding how quickly physiological responses are evolving over time.