WASHINGTON UNIVERSITY IN ST. LOUIS

McKelvey School of Engineering
Department of Computer Science & Engineering

Thesis Examination Committee:
Alvitta Ottley, Chair
Ganesh Babulal
Nathan Jacobs

Toward Vehicle-Agnostic Driving Signatures for Cognitive Impairment Prediction from Naturalistic Driving Data

by
Aadarsha Gopala Reddy

A thesis presented to
the McKelvey School of Engineering
of Washington University in
partial fulfillment of the
requirements for the degree
of Master of Science

May 2026
St. Louis, Missouri

Table of Contents

List of Figures

List of Tables

List of Abbreviations

AD Alzheimer's Disease

ADAS Advanced Driver-Assistance Systems

APOE Apolipoprotein E

AUC Area Under the Curve

CDR® Clinical Dementia Rating

DANN Domain-Adversarial Neural Network

GPS Global Positioning System

GRU-DANN Gated Recurrent Unit Domain-Adversarial Neural Network

LOGO Leave-One-Group-Out (participant-wise in this thesis)

MCI Mild Cognitive Impairment

PET Positron Emission Tomography

PR Precision-Recall

ROC Receiver Operating Characteristic

RPM Revolutions Per Minute

VIF Variance Inflation Factor

Acknowledgments

I would like to thank my supervisor, Dr. Ganesh Babulal, for his guidance and support throughout this research. I am also grateful to my committee members, Dr. Alvitta Ottley and Dr. Nathan Jacobs, for their valuable time, support, and feedback. I also thank Dr. David Brown for his consistent support and encouragement throughout this work, and Dr. Subrata Pal for his assistance.

Aadarsha Gopala Reddy

Washington University in St. Louis
May 2026

ABSTRACT OF THE THESIS

Toward Vehicle-Agnostic Driving Signatures for Cognitive Impairment Prediction from Naturalistic Driving Data

by

Aadarsha Gopala Reddy

Master of Science in Computer Science

Washington University in St. Louis, 2026

Professor Alvitta Ottley, Chair

This thesis studies whether naturalistic driving data can help predict binary Clinical Dementia Rating (CDR®) status while accounting for differences across vehicles. The final analytic dataset comprised 26,968 participant-weeks from 304 participants. Weekly driving features were derived from real-world telematics data and combined with four demographic covariates. Primary model comparisons used leave-one-participant-out (LOGO) cross-validation, with one individual held out at a time and pooled participant-level metrics used as the main reporting surface.

The main comparison includes six model families evaluated on the same dataset under a shared LOGO framework. Performance remained modest overall. GRU-DANN had the highest participant-level ROC AUC (0.599) and balanced accuracy (0.584), Logistic Regression had the highest sensitivity (0.523), and Random Forest was the strongest baseline on ranking-oriented metrics, with ROC AUC of 0.595 and PR AUC of 0.355. DANN did not clearly outperform simpler baselines.

Taken together, these results might suggest that naturalistic driving data may carry limited information related to cognitive status, but the present pipeline should be interpreted as an empirical comparison of modeling choices rather than as a clinical screening tool. The main

contribution of this thesis is a comparison of baseline, domain-adversarial, and sequence-based modeling choices, together with a clearer account of what this dataset and pipeline can and cannot currently support.

Chapter 1

Introduction

Alzheimer's Disease (AD) remains one of the most urgent global health challenges of the 21st century, affecting an estimated 55 million people worldwide, with projections suggesting that this number could reach 139 million by 2050 [8]. A major clinical problem is the long preclinical period, during which pathological brain changes are already underway but overt cognitive symptoms may still be limited or absent. That window matters because earlier identification could support monitoring, risk stratification, and eventual intervention. Yet the most established biomarker pathways, including cerebrospinal fluid analysis and positron emission tomography (PET), remain expensive, invasive, and difficult to scale for broad screening [18].

This gap has motivated growing interest in passively collected behavioral measures from everyday life. Naturalistic driving is especially relevant because it draws on navigation, motor control, executive function, attention, and decision-making during a routine real-world activity. Within the DRIVES (Driving Real-World In-Vehicle Evaluation System) project at Washington University in St. Louis, prior studies reported encouraging discrimination of AD-related outcomes from driving-derived variables, with area under the curve (AUC) values ranging from 0.64 to 0.90 in specific study settings [3, 5]. Those studies suggest that driving behavior may carry clinically relevant information. They do not, on their own, establish a clinically usable marker, and they do not remove the need for careful validation under stricter held-out participant and cross-vehicle conditions.

A central obstacle is vehicle heterogeneity. Drivers in a naturalistic cohort do not use standardized vehicles. They drive different makes, models, vintages, powertrains, and technology packages, often with different levels of Advanced Driver-Assistance Systems (ADAS). Those differences can change how similar underlying driving behavior appears in the telematics

stream, which makes it harder to separate driver-linked variation from vehicle-linked variation [10]. If a model learns vehicle-linked variation instead of behavior linked to the driver, then apparent predictive performance may not transfer cleanly across the broader fleet.

This thesis addresses that problem directly. It frames vehicle heterogeneity as a domain shift problem and asks whether domain adaptation can reduce vehicle-linked confounding while preserving information relevant to the driver's cognitive status. Domain-adversarial neural networks (DANNs) are used here as one test case because they are designed to learn representations that are useful for the main label while being less predictive of the nuisance domain [11, 12]. Chapter 1 therefore motivates the question: whether more transportable driving representations can be learned in this setting, and how that possibility should be evaluated empirically.

1.1 The Alzheimer's Disease Challenge

AD is a progressive neurodegenerative disorder and the leading cause of dementia, accounting for approximately 60% to 80% of cases worldwide [8]. The disease is associated with amyloid-beta plaques, tau pathology, neurodegeneration, and gradual cognitive decline. As populations age, the burden of AD is expected to rise sharply, placing increasing strain on patients, caregivers, and health systems.

The preclinical stage of AD is especially important because biological changes may be present long before clear functional decline is recognized. That stage can last for years, which creates an opportunity for earlier monitoring and for eventual intervention before more advanced impairment develops. The CDR® scale is commonly used to stage cognitive impairment, with CDR® 0 indicating no cognitive impairment, CDR® 0.5 indicating very mild impairment, and higher values indicating greater severity [14]. In practice, however, identifying at-risk individuals early enough for scalable follow-up remains difficult.

Current biomarker pathways are informative but hard to deploy widely. Cerebrospinal fluid analysis requires lumbar puncture, and PET imaging is costly, specialized, and not practical for large-scale community screening [18]. These constraints limit how often such measures can be used outside research or specialty settings. The result is a persistent gap between

scientific knowledge about early AD pathology and the practical ability to screen for risk in everyday clinical contexts.

That gap has helped drive interest in passive, ecologically valid behavioral markers. A useful passive behavioral measure would need to be low burden, scalable, and sensitive enough to justify further clinical assessment. Everyday driving is a plausible candidate because it is frequent, cognitively demanding, and measurable through unobtrusive sensing.

1.2 Naturalistic Driving as a Behavioral Signal for Cognitive Status

Naturalistic driving provides access to a complex real-world behavior that depends on multiple cognitive domains relevant to aging and dementia. Driving involves visual processing, spatial navigation, motor coordination, executive control, divided attention, and rapid decision-making. Unlike clinic-based tasks, it also unfolds in ordinary environments where people encounter familiar routes, traffic variation, and competing demands [3, 5].

The DRIVES project has built a substantial research program around this idea by collecting high-frequency sensor data from older adults during everyday driving. Participants' personal vehicles are equipped with commercial dataloggers that record accelerometer and gyroscope signals at 24 Hz, alongside Global Positioning System (GPS) coordinates and speed at 1 Hz. This produces dense longitudinal records of routine driving behavior, including acceleration, braking, turning, and trip-level movement patterns [3, 5, 7].

Prior DRIVES studies provide important motivation for the present work. Babulal et al. (2021) reported AUC values from 0.64 to 0.82 for driving variables alone in classifying preclinical AD biomarker positivity [3]. They also reported stronger performance after adding age, apolipoprotein E (APOE) $\varepsilon 4$ status, and neuropsychological measures. Bayat et al. (2023) reported related associations between everyday driving and plasma AD biomarkers, with F1 scores of 0.75 from driving variables alone and 0.80 after adding age and APOE $\varepsilon 4$ carrier status [5]. Recent DRIVES-related studies have further expanded the dataset and linked driving behavior with broader clinical, environmental, biomarker, and demographic context [1, 7].

These prior studies show that naturalistic driving is scientifically promising. Even so, they do not remove the need to test generalization under stricter deployment conditions. A signal that seems useful in one study design may weaken when transferred to new drivers or different vehicles. That concern matters here because the same maneuver can yield different telematics patterns across vehicles.

Within that broader search for practical behavioral markers, the promise of naturalistic driving depends on separating driver-linked behavior from context-linked variation.

1.3 The Vehicle Heterogeneity Problem

Vehicle heterogeneity refers to systematic differences across vehicles that can change measured driving signals even when the underlying driver behavior is similar. Those differences include physical properties, control systems, sensing context, and automation features. In a machine learning pipeline, such variability can act as a confound, making it difficult to tell whether a feature reflects the driver, the vehicle, or an interaction between the two.

Several sources of heterogeneity are especially relevant. First, vehicle mass and handling characteristics affect kinematic measurements such as lateral acceleration, braking dynamics, and turning smoothness. A heavier sport utility vehicle and a compact sedan can produce different accelerometer or gyroscope patterns during a similar maneuver. If those patterns enter a model without adequate control, the model may partially learn vehicle physics rather than behavior linked to cognitive status.

Second, ADAS features can directly alter the signals that many driving models treat as meaningful. Automatic emergency braking, adaptive cruise control, lane-keeping assist, and related systems may change deceleration patterns, steering variability, or lane-position behavior [10]. In those cases, an observed event in the telematics stream may reflect assistance logic as much as driver intent.

Third, powertrain differences matter. Electric vehicles and internal combustion vehicles can produce different acceleration profiles because of differences in torque delivery, drivetrain response, and regenerative braking behavior. As a result, telematics features that appear behaviorally meaningful in one vehicle class may not carry the same interpretation in another.

Fourth, sensor placement and calibration can introduce additional variability. Even with a common datalogger platform, placement within the vehicle and alignment relative to the vehicle axes can affect the measured magnitude and direction of motion signals.

Taken together, these issues create a domain shift problem. A model trained on one subset of vehicles may perform adequately within that subset while failing to transfer when vehicle-linked structure changes. For a driving-based cognitive biomarker, that is not a minor technical nuisance. It is a core threat to external validity and clinical usefulness.

1.4 Domain Adaptation as a Candidate Approach

Domain adaptation studies how to learn from one distribution while improving transfer to another related distribution. In this thesis, the motivating question is whether that framework can help when vehicle-linked variation shifts the distribution of driving features. The domain is therefore tied to vehicle-related context, while the prediction target remains cognitive status.

The theoretical basis for domain adaptation includes work by Ben-David et al. (2010), which formalized how target-domain error depends in part on source error and divergence between source and target distributions [6]. Later deep-learning approaches embedded that idea into representation learning, including Deep Adaptation Networks and adversarial methods that try to reduce domain-specific information in learned features [13].

Among those methods, DANN is a useful candidate for this thesis because it explicitly combines supervised prediction with adversarial pressure against domain identification [11, 12]. The architecture couples a feature extractor with two downstream objectives: a label predictor for the main task and a domain classifier for the nuisance domain. A gradient reversal layer allows the feature extractor to receive an adversarial signal from the domain classifier during backpropagation. In principle, that setup can encourage representations that remain useful for the label while carrying less recoverable information about the domain.

That principle is especially appealing for naturalistic driving because the central challenge is not only prediction, but prediction that transfers across heterogeneous vehicles. In that

setting, domain-adversarial modeling offers a concrete way to test whether learned representations can become less vehicle-dependent while remaining informative for cognitive status. This thesis therefore evaluates DANN against non-adversarial alternatives under its primary evaluation setting.

That evaluation setting is leave-one-participant-out cross-validation, corresponding to the standard leave-one-group-out (LOGO) protocol because each study participant defines one group in the main cross-model comparison. GroupShuffleSplit cross-validation is used for within-family ablation work, but cross-model claims require LOGO because the protocol holds out each individual exactly once under a common comparison design.

1.5 Research Questions and Study Objectives

This thesis asks whether naturalistic driving data can support clinically meaningful CDR® classification while accounting for the vehicle heterogeneity that complicates real-world deployment. The study is organized around three research questions and corresponding study objectives.

Research Question 1: Under LOGO, how much discriminative signal for binary CDR® classification is present in the available naturalistic driving representations?

Research Question 2: When vehicle-related variation is treated as a domain shift problem, does domain-adversarial modeling change classification behavior relative to non-adversarial baselines, and if so, in what direction?

Research Question 3: Which combinations of driving features, demographic covariates, vehicle context, and model design appear most promising for future work on more vehicle-robust driving biomarkers?

These questions lead to three study objectives. First, the thesis builds an end-to-end modeling pipeline that links weekly driving-derived representations with participant-level clinical labels, demographic covariates, and vehicle context. Second, it evaluates baseline machine learning models alongside domain-adversarial variants under a common LOGO framework, which provides the primary basis for comparison. Third, it uses those results to clarify

what the current data and modeling choices can support, and where the main limits to generalization still remain.

1.6 Contributions

This thesis makes technical, clinical, and data contributions to the study of passive behavioral measures for cognitive decline.

Technical Contribution: The thesis formulates vehicle heterogeneity as a modeling problem that must be confronted directly rather than treated as background noise. It implements and compares multiple model families, including domain-adversarial and non-adversarial approaches, within a common evaluation pipeline centered on LOGO. It also develops a feature-engineering and data-integration workflow that combines driving-derived summaries with selected demographic and vehicle context without assuming that any single representation has already solved the generalization problem.

Clinical Contribution: The work sharpens the translational question facing driving-based biomarkers for AD. Instead of asking only whether driving data can predict cognitive status in a favorable setting, it asks whether such prediction remains credible when vehicle-linked variation is taken seriously. That framing is clinically important because any eventual screening approach would need to operate across a heterogeneous community fleet, not a standardized fleet.

Data Contribution: The thesis develops procedures for aligning longitudinal driving data with intermittent clinical assessments and for incorporating vehicle-level characteristics into the analytic dataset. Those steps create a reusable structure for studying how behavioral measurements, participant context, and vehicle context interact in longitudinal observational data.

Together, these contributions position the thesis as an empirical test of what is currently feasible, what remains uncertain, and what kinds of methodological changes may be needed before naturalistic driving can serve as a more transportable behavioral marker of cognitive status.

1.7 Thesis Organization

The remainder of this thesis follows the usual sequence of related work, methods, results, and discussion. Chapter 2 reviews prior literature on driving, AD, passively collected measures from everyday life, and domain adaptation. Chapter 3 describes the dataset, preprocessing, feature construction, modeling pipeline, and evaluation design. Chapter 4 presents the main empirical results. Chapter 5 interprets those results, discusses limitations, and outlines next steps. Chapter 1 has a narrower role: define the problem, motivate the study, and set expectations for the analyses that follow.

Chapter 2

Literature Review

2.1 Preclinical Alzheimer's Disease and the Need for Scalable Monitoring

AD remains the leading cause of dementia and is associated with a long preclinical interval during which pathological changes may accumulate before overt functional decline is recognized [8]. That interval creates a practical monitoring problem. For early detection, useful markers must be informative before dementia is clinically obvious and feasible to repeat over time. The CDR®remains important for staging cognitive and functional impairment in both research and clinical settings, but it does not by itself solve the challenge of scalable longitudinal monitoring [14].

Recent biomarker reviews clarify that the limitation is not biological relevance, but deployability. Fluid and imaging biomarkers provide valuable diagnostic and prognostic information, yet they remain comparatively costly, invasive, specialized, or difficult to repeat broadly in community settings [4]. Translational work therefore increasingly frames early detection as a complementary multimodal problem: fluid and imaging markers remain central, but lower-burden measures may be needed when the goal is repeated observation, triage, or broader screening rather than one-time confirmation alone [4].

2.2 Everyday Sensing and Passive Measures

Passively collected measures sit within that lower-burden monitoring landscape because they use digital devices to capture behavior, physiology, or cognition outside clinic visits. In preclinical AD, remote and unsupervised assessments are attractive because they can sample

routine behavior more frequently and in more natural settings than occasional in-person testing [15, 16]. This advantage is especially relevant when subtle decline may only be visible over repeated observations rather than in sparse clinical tests.

The same literature also shows why these passively collected measures remain methodologically fragile. Reviews of remote cognitive assessment and broader AD sensing measures consistently note uneven validation, limited external replication, weak calibration reporting, and uncertain generalizability across settings and devices [15, 16, 4]. The field is therefore not only constrained by whether everyday data contain signal, but also by whether the extracted signal is stable, interpretable, and reliable enough to support future clinical use.

2.3 Naturalistic Driving as a Cognitive Biomarker

Within the wider landscape of passively collected measures, naturalistic driving is compelling because it is both common and cognitively demanding. Driving requires attention, visuospatial processing, motor coordination, planning, and rapid decision-making in an ecologically valid setting, making it a plausible behavioral marker of early cognitive change. The strongest case for driving, however, comes from the progression of empirical studies that have tested different outcomes, feature sets, and analytic strategies.

Early work from the DRIVES Project provided proof-of-concept evidence that everyday driving variables can help distinguish biomarker-defined preclinical AD in cognitively normal older adults [3]. Later work extended this line of inquiry by linking naturalistic driving behavior with plasma amyloid biomarkers, suggesting that lower-burden biological and behavioral signals can be studied together [5]. More recent studies expanded the modeling context by incorporating environmental, socioeconomic, and genetic information alongside driving features [1]. Newer deep-learning work explored naturalistic driving for early mild cognitive impairment detection in a much smaller cohort [2]. In parallel, the DRIVES data resource has matured into a broader multimodal platform that links naturalistic driving with clinical and neurobehavioral measurements [7]. Table 2.1 summarizes the main empirical naturalistic-driving studies that most directly motivate the modeling questions taken up later in this thesis.

Table 2.1: Selected empirical naturalistic-driving studies relevant to the Chapter 2 literature review. Each study is organized using a fixed comparison template: population, task, signal/input, main finding, thesis relevance, and limitation.

Item	Value / note
Babulal et al. (2021) [3]	
Population	131 cognitively normal older adults.
Task	Classify biomarker-defined preclinical AD using naturalistic driving behavior.
Signal/input	Six in vivo driving variables, with follow-on models adding age, APOE $\varepsilon 4$, and neuropsychological measures.
Main finding	Driving variables alone yielded AUC values from 0.64 to 0.82, improving to 0.81–0.90 when additional covariates were added.
Thesis relevance	Establishes proof of concept for naturalistic driving as an early cognitive biomarker.
Limitation	The study is not framed around vehicle heterogeneity, and the authors note limited generalizability of the sample.
Bayat et al. (2023) [5]	
Population	142 cognitively normal older adults.
Task	Predict plasma amyloid positivity from naturalistic driving features.
Signal/input	Six driving features in an ANN, with a second model adding age and APOE $\varepsilon 4$ status.
Main finding	The driving-only model achieved an F1 score of 0.75, improving to 0.80 after adding age and APOE information.
Thesis relevance	Extends the driving literature toward blood-based biomarkers.
Limitation	The study does not explicitly address cross-vehicle generalization or domain shift.
Al-Hammadi et al. (2025) [1]	
Population	292 participants and 2,792 observations.
Task	Classify preclinical AD using multimodal predictors.
Signal/input	Naturalistic driving, cognitive screening, neighborhood deprivation, sociodemographic factors, and genetic risk variables.

Continued on next page

Item	Value / note
Main finding	Adding environmental and socioeconomic context modestly improved predictive performance, and model accuracy was quantified in a held-out 20% test set across 100 resampling rounds.
Thesis relevance	Shows that driving features can be evaluated within a richer contextual model.
Limitation	The primary heterogeneity focus is socioeconomic and environmental rather than vehicle-linked variation.
Al-Hindawi et al. (2026) [2]	
Population	19 final participants. 7 with mild cognitive impairment [MCI] and 12 normal.
Task	Early mild cognitive impairment detection.
Signal/input	Deep-learning models trained on driving traces collected from participants' personal vehicles under naturalistic conditions.
Main finding	Full-trip representations outperformed turn-only inputs, with the best-performing model reported at 78%.
Thesis relevance	Demonstrates ongoing interest in deep-learning approaches for naturalistic driving.
Limitation	The cohort is small and does not provide evidence that domain-adversarial methods resolve vehicle-linked shift.

Taken together, these studies support a cautious conclusion. Naturalistic driving appears informative enough to justify serious modeling work, but the evidence base remains heterogeneous in outcome definitions, modeling targets, and validation choices. The literature is stronger at showing that driving carries signal than at showing that the signal is already robust to all of the contextual differences encountered in real-world fleets.

Naturalistic driving research also draws on a wider telematics and trajectory-analysis literature that explains how raw sensor streams can be converted into interpretable summaries such as speed variability, jerk, braking intensity, route regularity, and trip timing [20, 17]. That methodological literature matters because it clarifies how behavioral features are built, but it also highlights an important limit: feature extraction alone does not guarantee construct validity. A useful telematics feature must capture something stable about the driver or behavior of interest rather than merely reflecting quirks of sensing, context, or platform.

2.4 Vehicle Heterogeneity as a Confounder

One reason this distinction matters is that naturalistic driving cohorts rarely rely on a standardized fleet. Participants drive vehicles with different weights, powertrains, chassis properties, sensing environments, and driver assistance systems. Reviews of driving heterogeneity argue that such variation should be treated as structured behavioral and measurement diversity rather than as random noise to be averaged away [19]. For driving-based cognitive modeling, this means the same observed telematics feature can encode both driver behavior and the vehicle through which that behavior is expressed.

Several mechanisms make that confounding plausible. Physical vehicle properties can alter acceleration, braking, cornering, and ride dynamics, so the same maneuver may yield different motion signatures across platforms. ADAS features add another layer because systems such as adaptive cruise control, lane support, or automatic emergency braking can change steering or deceleration patterns in ways that are behaviorally meaningful for safety but difficult to attribute cleanly to the driver [10]. Powertrain differences, including electric versus internal-combustion propulsion, can also shift observed distributions through different torque delivery and braking dynamics.

The literature therefore supports treating vehicle heterogeneity as a real interpretive issue. If observed features are partly vehicle-linked, then a model may appear accurate while relying on regularities that will not transfer cleanly across the broader fleet. That possibility is not yet the same as proving failure, but it is enough to make generalizability a central methodological issue rather than a secondary nuisance term [19, 10].

2.5 Domain Adaptation for Heterogeneous Behavioral Data

Domain adaptation is relevant when training data and deployment data come from related but nonidentical distributions. In that setting, target-domain performance depends not only on how well a model fits the source data, but also on how different the domains are and on whether a shared hypothesis can perform well across them [6]. This basic logic makes domain

adaptation a natural conceptual response whenever the observed data can shift because of context, platform, or acquisition differences.

Later deep-learning methods translated that idea into concrete representation-learning strategies. Deep Adaptation Networks explicitly reduce mismatch through distribution matching in learned feature spaces [13]. DANNs then advanced a more direct objective: learn features that remain useful for the main task while becoming less informative about domain membership through adversarial training and gradient reversal [11, 12]. These methods do not guarantee success in every application, but they provide a principled way to test whether prediction is relying too heavily on domain-specific structure.

The broader sensing and health literature reinforces why this matters. Reviews of transfer learning in digital health emphasize that sensing data are fragmented, heterogeneous, and often collected under shifting device or user conditions [9]. Across these settings, the recurring lesson is not that domain adaptation always succeeds, but that apparent performance in one measurement context may say little about performance in another if distribution shift is ignored. That literature does not prove that domain-adversarial training will solve vehicle heterogeneity in naturalistic driving, but it does justify evaluating it as a response to context-linked shift rather than treating it as an arbitrary model choice.

2.6 Summary and Gap Statement

The literature reviewed in this chapter supports three broad conclusions. First, preclinical AD creates a real need for monitoring tools that are lower burden and more scalable than conventional biomarkers alone [8, 4]. Second, passively collected measures are promising partly because they enable repeated observation outside the clinic, but their translational value is still limited by weak external validation, uncertain transportability, and uneven clinical implementation [15, 16, 4]. Third, naturalistic driving has produced enough encouraging results to motivate continued modeling work, yet those results do not by themselves show that learned signals are already robust to vehicle-linked variation [3, 5, 1, 2, 19, 10].

What remains unresolved is therefore not whether driving data can be modeled in a way that remains credible when the sensing context changes across real vehicles. The gap motivating the rest of this thesis is a methodological one: how to construct a workflow that

links longitudinal driving summaries to clinically meaningful labels, acknowledges vehicle-linked heterogeneity as a possible source of domain shift, and compares adversarial and non-adversarial models under a common held-out participant protocol. Chapter 3 turns to that operational problem by describing the dataset construction, feature generation, model design, and evaluation framework used in this thesis.

Chapter 3

Methods

3.1 Study Overview and Data Sources

This study used a multi-stage data-processing workflow that combined naturalistic driving telemetry, study linkage data, demographic and cognitive assessment records, and vehicle-derived domain labels. Table 3.1 summarizes the stepwise cohort retention and the main filtering logic at each stage, while Table 3.2 provides the resulting dataset totals, class composition, demographics, and analysis units that anchor the rest of this chapter.

In brief, Stage 1 parsed the Azuga high-frequency trip archive into synchronized trip-level sensor tables spanning accelerometer, gyroscope, GPS, vehicle speed, and engine revolutions per minute (RPM). Stage 2 linked those trip features to the study metrics export, preserving participant identifiers, device-linkage fields, calendar-based grouping variables, and trip-context measures such as time of day, overspeeding, idling, stop time, and trip-length categories. Stage 3 aggregated retained trips into participant-week rows, attached the temporally closest usable CDR®assessment, and produced the weekly modeling dataset. The primary analytic unit is therefore the participant-week, whereas final cross-model comparison is carried out at the participant level. For the DANN and GRU-DANN, the adversarial domain labels come from a four-cluster K-means surface derived from vehicle-specification features. Figure 3.1 provides a compact visual summary of this three-stage workflow.

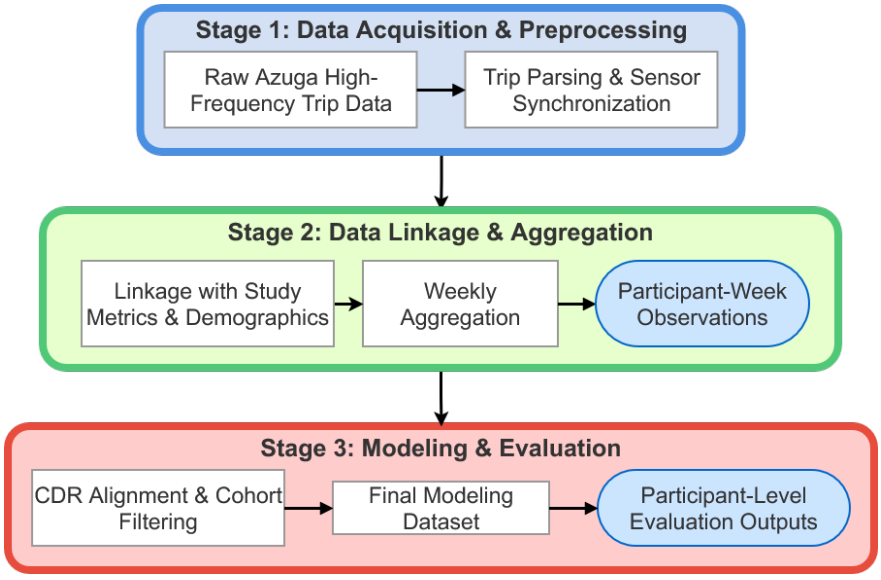

Figure 3.1: Three-stage workflow used in this thesis, from raw Azuga trip telemetry through study linkage and weekly aggregation to final modeling and participant-level evaluation outputs.

Table 3.1: Stepwise cohort retention and filtering logic for the Chapter 3 pipeline. The table reports verified stage-level counts and the main exclusion logic at each stage; it does not imply separately tabulated losses for every individual Stage 3 filter.

Stage	Retained cohort	Main inclusion / exclusion logic
1: telemetry archive	790,560 trips from 429 vehicles	Raw Azuga trip files were parsed into synchronized trip-level sensor tables. Trips with no usable trajectory or speed information, or with too few valid records for downstream feature extraction, were not retained in the clean Stage 1 output.
2: linked trip table	645,610 trip rows from 419 vehicles and 350 participants	Duplicate trip records were removed before linkage, and only trips that matched across the telemetry and study-metrics sources were retained. This reconciliation step reduced the vehicle count from 429 to 419.

Continued on next page

Stage	Retained cohort	Main inclusion / exclusion logic
3: analytic weekly dataset	26,968 participant-weeks from 304 participants	Trip-level quality filters removed very short, implausible, or high-missingness trips; participants with less than 365 days of retained driving coverage were excluded; weekly rows were then labeled by the nearest CDR®assessment and retained only when the assessment fell within ±365 days of the weekly anchor.

Table 3.2: Dataset lineage, cohort composition, labels, demographics, domain composition, and analysis units for the final analytic dataset used in Chapter 3.

Item	Value / note
Dataset lineage	
Final dataset	26,968 participant-weeks from 304 participants
Calendar span	2022-04-18 through 2024-04-15
Reconciliation loss	10 vehicles lost during reconciliation
Post-linkage attrition	Participant count narrowed from 350 to 304 after aggregation and cohort filtering
Outcome and labels	
Healthy class definition	CDR®0.0
Impaired class definition	CDR®$\geq$ 0.5
Weekly-row class balance	20,239 healthy vs. 6,729 impaired 75.0% vs. 25.0%
Participant-level class balance	218 healthy vs. 86 impaired 71.7% vs. 28.3%
CDR®severity snapshot	CDR®0.0: 232 participants CDR®0.5: 70 participants CDR®1.0: 2 participants

Continued on next page

Item	Value / note
CDR®timing offset	Median 96 days, Mean 101.6 days
	Q1 48, Q3 144, range 0–365
Demographics	
Age at CDR®assessment	Median 74 years, Mean 74.2 years
	Q1 70, Q3 78, range 54–92
Education	Median 16 years, Mean 16.5 years
	Q1 15, Q3 18, range 12–24
Gender study codes	1: 134 participants
	2: 170 participants
Race study codes	4: 42 participants
	5: 262 participants
Evaluation units and vehicle-domain grouping	
Weeks per participant	Median 96, Mean 88.7
	Q1 80, Q3 102, range 4–105
Trips per participant-week	Median 20, Mean 21.8
	Q1 12, Q3 29, range 1–113
Final reporting unit	304 held-out participant predictions under LOGO
Missingness after cleaning	0% missingness for retained demographics, `cdr_label`, `vehicle_cluster`, and `n_trips`
K=4 vehicle counts	Cluster 0: 64 vehicles
	Cluster 1: 74 vehicles
	Cluster 2: 163 vehicles
	Cluster 3: 127 vehicles
Cluster-use note	These vehicle counts describe the refreshed vehicle-domain grouping used for domain labels in the main comparison

3.2 Data Processing Pipeline

The first processing stage operates at the trip level. Each raw Azuga trip file is parsed into a chronological sensor table containing accelerometer, gyroscope, GPS, vehicle-speed, and engine-RPM information. Timestamps are converted to Coordinated Universal Time (UTC), speed is harmonized by preferring on-board vehicle readings when available and otherwise falling back to GPS-derived speed, and all downstream event calculations are based on observed time gaps between successive records rather than on an assumed constant interval. From that time-aligned trip table, the pipeline derives speed summaries, acceleration and braking events, jerk-based features, turning measures, trip duration and approximate distance, geographic start and end anchors, and data-quality indicators such as per-sensor missingness and irregular time gaps. Trips with no usable speed information or too few valid records are removed before downstream analysis. Processing was parallelized across multiple CPU cores, and the clean trip table was written in a format suitable for reproducible downstream use.

The merge stage aligns those trip-level features with the study metrics dataset. Duplicate trip records are removed before joining, and only trips that can be matched across the telemetry and study datasets are retained. The merged table then carries forward the extracted driving features together with participant identifiers, device-linkage information, calendar-based grouping variables, and trip-context variables such as time-of-day indicators, overspeeding measures, idling time, stop time, and trip-length categories.

The weekly aggregation and labeling stage converts the merged trip table into the modeling dataset. It first applies trip-level exclusion rules to remove implausible or uninformative trips: duration shorter than 60 seconds, distance shorter than 0.1 miles, maximum speed above 120 mph, duration longer than 180 minutes, or more than 50% missing speed data. Sparse sensor-derived jerk features are then set to zero when the absence of a signal indicates that no qualifying event was observed rather than a failed measurement. Vehicle-domain assignments are attached at the trip level before aggregation so that each weekly observation reflects the vehicle context actually used during that week. After that, a participant-level filter removes participants whose span between first and last retained trip is less than 365 days. Relative to the 429 vehicles present at stage 1, 10 vehicles (2.3%) were lost during telemetry-to-study reconciliation, and the participant count then narrowed from 350 after linkage to 304 after weekly aggregation and cohort filtering.

Weekly aggregation is then carried out at the participant-week level using ISO calendar weeks. Count and duration variables are summed across trips, normalized rate or quality measures are averaged, binary indicators are reduced to whether they occurred at least once during the week, and geographic anchors are taken from the earliest and latest valid trip in the week. The resulting weekly record summarizes both how often events occurred and what a typical trip looked like during that observation window. It also retains trip counts, the week start date, the modal vehicle assignment for that week, and the weekly domain label. Each row in the final modeling table is therefore one weekly observation for one participant, not one prediction for a whole participant. That distinction matters later because the tabular models and the sequence model reach participant-level outputs in different ways.

3.3 Feature Construction

Feature construction began at the trip level and continued after weekly aggregation. The trip-level extraction stage computes speed statistics, hard braking and sudden acceleration events, threshold-based acceleration and braking counts, jerk summaries from speed, RPM, and accelerometer channels, gyroscope-derived turning measures, missing-data percentages, geographic anchors, route-complexity features, and mid-trip stop measures. Route-complexity features include heading-change summaries, heading entropy, displacement ratio, turn counts at multiple angle thresholds, and a composite route-complexity index. Automated checks were used to verify the underlying geometric calculations, threshold counting, and stop-detection logic.

At the weekly level, a fixed aggregation scheme is applied and then supplemented with exponentially weighted moving variance (`ewma_var_*`) and standard deviation (`ewma_sd_*`) features for selected week-to-week behavioral changes. After exclusion of identifiers, label fields, weekly metadata, domain targets, raw demographics, and vehicle-specification columns, the initial behavioral pool contained 116 numeric features. Appendix A provides the complete variable-by-variable feature dictionary. Vehicle-specification and ADAS fields are kept separate from this pool because they are used to derive vehicle-domain labels rather than to define the main behavioral representation used for model comparison.

For the main comparison reported in the thesis, the behavioral representation was fixed in advance to the 84-feature subset labeled `no_gps_no_ewma` in the codebase. This version

removes four raw GPS anchor coordinates (`first_lat`, `first_lon`, `last_lat`, and `last_lon`) together with 28 exponentially weighted moving variance or standard-deviation features. The intent was to retain variables that describe driving behavior while excluding raw location anchors and a large family of highly redundant volatility summaries.

That choice was motivated by both empirical and structural considerations. In earlier LOGO feature-set ablations, the strongest repeatable gains appeared after the EWMA family was removed, and the `no_gps_no_ewma` condition performed best overall in that historical screening exercise across the six model families. The structural rationale was similar: `ewma_var_*` and `ewma_sd_*` pairs are mathematically linked because standard deviation is the square root of variance, and several EWMA-derived fields also showed severe multicollinearity. For example, `ewma_var_d_hard_braking` and `ewma_var_d_braking_count_8` had variance inflation factor (VIF) values above 60, while multiple duration- and distance-related EWMA fields behaved similarly. The GPS-coordinate removal is interpreted more narrowly: it keeps the primary behavioral representation centered on behavior rather than on raw start and end location anchors. These earlier ablation results are used here only to justify freezing the feature representation before the final comparison, not as evidence for the final cross-model ranking.

Demographic features are handled separately from the behavioral features. The retained demographic set contains four variables: age at CDR®assessment, education, gender, and race in encoded form. Table 3.2 summarizes the cohort composition for those variables and reports the observed study-code frequencies for gender and race as they appear in the analytic dataset. The repository documents the encoded covariates used for analysis and their reproducible encoding step, but it does not include the authoritative production codebook needed to assign human-readable semantics to raw study codes 1/2 or 4/5. Accordingly, these codes are reported here as observed counts only, without semantic interpretation.

3.4 Outcome Definition and Cohort Filtering

The prediction target is a binary CDR®label derived during the cleaning and labeling stage. For each weekly participant row, the available CDR®records for that participant are examined, the absolute offset between the weekly anchor date and each CDR®date is computed, and the temporally closest usable assessment is identified. A weekly row is retained only if

the nearest assessment falls within a one-year window (±365 days) of that weekly anchor; rows beyond the window are dropped. This rule reflects a pragmatic tradeoff between temporal proximity and retention of a usable longitudinal sample in a cohort with intermittent clinical visits. The realized offsets in Table 3.2 show that the one-year rule functions as an upper bound rather than a typical match, with a median offset of 96 days and an interquartile range of 48 to 144 days.

The binary mapping is fixed in code: CDR®0.0 maps to 0, whereas CDR®0.5, 1.0, 2.0, and 3.0 each map to 1. In other words, CDR®0 remains the cognitively healthy class, and any CDR®value of 0.5 or higher is mapped to the impaired class. This mapping is hard-coded rather than estimated from the sample, which keeps the target definition stable across reruns. Table 3.2 reports the corresponding weekly-row and participant-level class balances, the participant-level CDR®severity snapshot, and the distribution of timing offsets between each weekly anchor and its selected CDR®assessment. Because the retained impaired group is dominated by CDR®0.5 (70 participants) with only two CDR®1.0 participants, the binary target should be interpreted primarily as distinguishing cognitively healthy participants from participants with predominantly very mild impairment.

Several cohort filters are therefore active before any model is trained. Trips failing the basic quality checks are excluded first. Participants with less than one year of retained driving coverage are removed next. After weekly aggregation, only rows that satisfy the CDR®alignment rule described above are kept. The final modeling dataset is therefore a filtered weekly cohort with participant identifiers for evaluation, vehicle-based domain labels for adversarial training, and one binary cognitive target per retained week. This design keeps the labeling rule explicit and bounded, but the temporal separation between some weekly anchors and their matched CDR®assessments should still be understood as a potential source of label noise when studying subtle cognitive change.

3.5 Modeling Approaches

The thesis comparison uses six model families under a common evaluation framework: Logistic Regression, XGBoost, Random Forest, multilayer perceptron (MLP), DANN, and Gated Recurrent Unit DANN (GRU-DANN). The first five models are easiest to understand as

weekly tabular models. They read one weekly observation at a time, represented as a fixed-length feature vector, and return a probability for that week. When demographic features are enabled for the baseline models, those four variables are appended directly to the same tabular input vector as the selected behavioral features. This is an early-fusion design.

DANN is also a weekly tabular model, but it adds an adversarial domain objective. Its architecture uses a multilayer perceptron feature extractor, a binary label predictor, and a domain classifier connected through a gradient reversal layer. In the implemented model, the weekly behavioral features are first mapped to a learned representation by the feature extractor. Demographic covariates do not pass through that feature extractor and do not enter the domain-classification branch. Instead, they are concatenated with the learned behavioral representation only at the label-prediction stage. This is a classifier-level fusion design rather than early fusion. The label predictor then maps that combined representation to a binary cognitive prediction, while the domain classifier is trained adversarially on the learned behavioral representation alone. Class imbalance is handled through weighted binary loss.

GRU-DANN is different in kind because it does not treat weeks as independent rows. Instead, it groups all retained weeks for one participant in chronological order and reads that ordered sequence as a short behavioral history. The GRU encoder reduces that sequence to one learned participant representation before a downstream multilayer perceptron stage. In the primary configuration, the sequence encoder uses two GRU layers with hidden size 64.

For GRU-DANN, the ordered weekly driving sequence is first encoded into a participant-level behavioral representation. Then, like DANN, GRU-DANN also uses a classifier-level fusion design. The tabular models first score weekly observations and later summarize those weekly scores, whereas GRU-DANN reads the ordered weeks together and produces one participant-level score directly. Figure 3.2 summarizes the shared domain-adversarial structure of DANN and the temporal extension used in GRU-DANN.

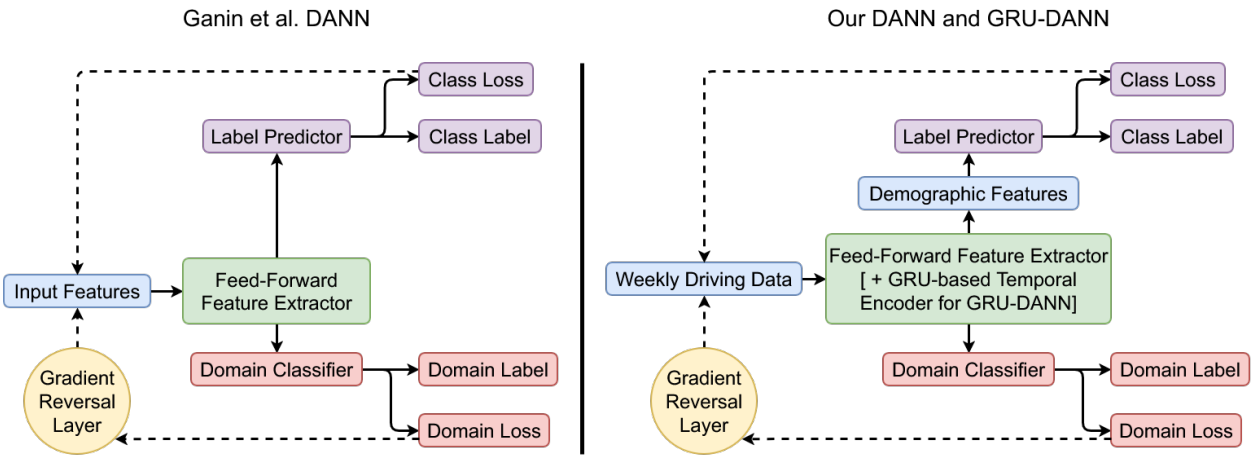

Figure 3.2: Conceptual comparison of the original DANN (left) and the DANN/GRU-DANN implementations used in this thesis (right). Both retain a feature extractor, label predictor, and domain classifier linked through gradient reversal. In both implementations, demographic features bypass the adversarial branch and are concatenated only before the label predictor; GRU-DANN additionally extends the behavioral-encoding stage with GRU-based temporal processing of ordered weekly driving histories. The base DANN concept follows Ganin et al. [12].

Table 3.3: Compact comparison-time configuration summary for the six thesis model families. All models used the fixed `no_gps_no_ewma` behavioral representation with demographics enabled, were evaluated under LOGO, and contributed participant-level outputs to the final comparison.

Model	Input / output unit	Demographic handling	Key comparison setting
Logistic Regression	Weekly tabular → participant summary	Early fusion	Participant-level output formed by aggregating weekly probabilities under LOGO.
XGBoost	Weekly tabular → participant summary	Early fusion	Same as other baselines.
Random Forest	Weekly tabular → participant summary	Early fusion	Same as other baselines.
MLPClassifier	Weekly tabular → participant summary	Early fusion	Same as other baselines.

Continued on next page

Model	Input / output unit	Demographic handling	Key comparison setting
DANN	Weekly tabular → participant summary	Late fusion before label predictor only	Domain-adversarial training with $\lambda_{\text{GRL}} = 0.01$; the domain classifier receives the learned behavioral representation only.
GRU-DANN	Ordered participant-week sequence → direct participant output	Late fusion before label predictor only	Two GRU layers, hidden size 64, and $\lambda_{\text{GRL}} = 0.01$.

Adversarial domain labels are taken from vehicle-based domain groupings rather than participant identity. In contrast, the outer cross-validation loop holds out one participant at a time. Participant identity therefore defines the evaluation split, whereas vehicle-domain grouping defines the nuisance domain that the DANN and GRU-DANN models are trained to deemphasize. Table 3.2 summarizes the vehicle, participant-week, and dominant-participant counts across clusters 0 through 3, and Table 3.3 summarizes the comparison-time configuration used for each model family. The baseline models do not use `vehicle_cluster` as an input feature.

3.6 Evaluation Protocol

Two evaluation protocols are used for different purposes. GroupShuffleSplit is used only for within-family ablation and hyperparameter search. Its role is to rank configurations within a model family, not to support thesis-level claims about one model family outperforming another. The final cross-model comparison is therefore carried out under leave-one-participant-out (LOGO) cross-validation. Under LOGO, each individual is held out exactly once, so all model families are evaluated on the same sequence of held-out cases even though the tabular models consume weekly observations inside each fold. Figure 3.3 summarizes this two-stage evaluation design.

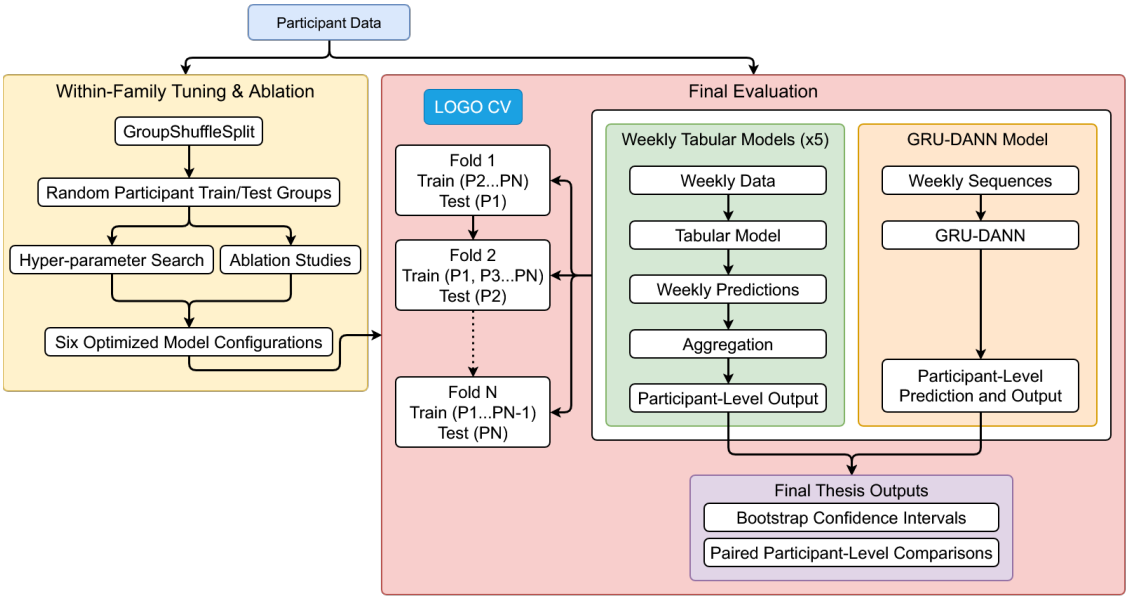

Figure 3.3: Evaluation design used in this thesis. GroupShuffleSplit is used only for within-family tuning and ablation, whereas leave-one-participant-out (LOGO) cross-validation is used for the final cross-model comparison. Weekly tabular models generate weekly probabilities that are averaged within each held-out individual and then thresholded at 0.5 to obtain participant-level classifications, while GRU-DANN produces participant-level outputs directly; bootstrap intervals and paired comparisons are then computed from the final participant-level predictions.

In the main comparison, each fold holds out one individual and trains on the remaining data. All weekly observations from that held-out individual are excluded from the training set for that fold. Missing-data handling occurs in two stages. Earlier in the pipeline, sparse RPM-jerk and accelerometer-jerk features are set to zero when the absence of a recorded event means that no qualifying event occurred, rather than that the measurement failed. During LOGO evaluation, any remaining missing numeric feature values are median-imputed within each fold using only the training data, and the held-out observations are transformed with that fitted imputer rather than contributing to it. Feature standardization follows the same rule: means and standard deviations are computed from the training fold only. Demographic covariates are required to be present after cleaning and are therefore not statistically imputed. In the final weekly dataset, missingness for the retained demographic fields, the binary CDR®label, `vehicle_cluster`, and the weekly trip-count field is 0% after cleaning.

All models contribute participant-level results to the final comparison. For the weekly tabular models, those outputs are obtained by taking the mean of the held-out weekly predicted

probabilities within each held-out individual and then applying a decision threshold of 0.5. Majority vote was not used in the final LOGO evaluation path. GRU-DANN, on the other hand, produces a participant-level output directly from the ordered weekly sequence. The weekly-row level refers to the 26,968 retained participant-week observations that enter feature construction and fold-specific scoring, whereas the participant-level evaluation refers to the 304 held-out individual predictions used for cross-model comparison.

This leads to two complementary output layers. Fold-wise results are retained as diagnostics. Participant-level results are the primary outputs when available because the final comparison is made at the participant level rather than the weekly level. To quantify uncertainty in those participant-level summaries, the evaluation uses 1,000 bootstrap replicates. For each model, the held-out participant predictions are resampled with replacement to create many alternative participant sets of the same size, and the participant-level metrics are recomputed on each resample. This bootstrap operates on the final participant-level predictions rather than retraining the models 1,000 times. More specifically, it asks how much participant-level AUC, balanced accuracy, sensitivity, and specificity would vary if an equivalent study cohort were sampled repeatedly.

The same participant-level predictions also support paired model-comparison outputs. For any two models that produced aligned held-out participant predictions for the same participants, the evaluator computes paired bootstrap differences by resampling the same participants for both models and recomputing the difference in participant-level metrics on each replicate. This yields a direct uncertainty interval for metric deltas such as the difference in ROC AUC, balanced accuracy, sensitivity, specificity, PR AUC, and Brier score. For ROC AUC specifically, the evaluator also records the paired nonparametric DeLong comparison on the same participant-level predictions. These pairwise outputs provide a supplementary view of uncertainty around the reported model differences, while the main comparison remains anchored to the participant-level summary metrics in Chapter 4.

These evaluation rules are kept fixed across the main comparison so that reported differences reflect model behavior rather than changes in feature handling, missing-data treatment, or participant-level reporting. Reproducibility is further supported by fixed random seeds, stable intermediate outputs, and automated validation checks on key feature-engineering and evaluation steps, which make the workflow rerunnable without changing the scientific protocol.

Chapter 4

Results

4.1 Participant-Level Comparison

Table 4.1 summarizes the participant-level results for the main six-model comparison. GRU-DANN has the highest ROC AUC (0.599) and balanced accuracy (0.584), Logistic Regression has the highest sensitivity (0.523), and Random Forest has the highest ROC AUC (0.595) and PR AUC (0.355) among the baseline models. DANN does not rank first on any reported metric. Because the participant-level impaired-class prevalence is 0.283, PR AUC values should be interpreted relative to that baseline rather than in isolation.

Table 4.1: Participant-level summary metrics for the selected six-model Chapter 4 comparison. ↑ indicates higher is better.

Model	ROC AUC ↑	95% CI	PR AUC ↑	Bal. Acc. ↑	Sensitivity ↑	Specificity ↑
Logistic Regression	0.586	[0.511, 0.652]	0.345	0.576	0.523	0.628
XGBoost	0.502	[0.426, 0.575]	0.292	0.506	0.209	0.803
Random Forest	0.595	[0.525, 0.661]	0.355	0.535	0.198	0.872
MLP	0.573	[0.503, 0.644]	0.337	0.527	0.186	0.867
DANN	0.554	[0.476, 0.626]	0.327	0.522	0.209	0.835
GRU-DANN	0.599	[0.525, 0.669]	0.346	0.584	0.512	0.656

Two observations are notable. First, performance remains modest overall: all models occupy a narrow participant-level ROC AUC range near 0.50 to 0.60. Second, no single model dominates across every metric. GRU-DANN leads ROC AUC and balanced accuracy, Logistic Regression leads sensitivity, and Random Forest is the strongest baseline on ranking-oriented metrics. The results are therefore better interpreted as a set of tradeoffs than as a single-model victory.

The 95% intervals in Table 4.1 should be interpreted as per-model participant-level bootstrap uncertainty for ROC AUC. In the main table, ROC AUC is used as the compact per-model uncertainty summary because it is the primary ranking metric for the overall comparison. For

PR AUC, balanced accuracy, sensitivity, and specificity, the uncertainty information is more informative when reported as paired-bootstrap contrasts on the same held-out participants. Table 4.2 therefore pulls a compact subset of those pairwise comparisons into the main text, while the full pairwise participant-level comparison surface remains available in Appendix B, Tables B.1 and B.2.

Table 4.2: Selected pairwise participant-level comparisons pulled forward from Appendix B. Differences are reported as model A minus model B on the same held-out participants.

Model pair	Selected paired results
Logistic Regression vs. XGBoost	
ΔROC AUC	0.084, 95% CI [0.008, 0.152], DeLong $p = 0.027$.
ΔPR AUC	0.053, 95% CI [-0.017, 0.119].
Δbalanced accuracy	0.070, 95% CI [-0.003, 0.135].
Δsensitivity	0.314, 95% CI [0.181, 0.423].
Δspecificity	-0.174, 95% CI [-0.245, -0.101].
Random Forest vs. GRU-DANN	
ΔROC AUC	-0.004, 95% CI [-0.085, 0.078], DeLong $p = 0.920$.
ΔPR AUC	0.009, 95% CI [-0.070, 0.091].
Δbalanced accuracy	-0.049, 95% CI [-0.118, 0.022].
Δsensitivity	-0.314, 95% CI [-0.435, -0.188].
Δspecificity	0.216, 95% CI [0.152, 0.284].
DANN vs. GRU-DANN	
ΔROC AUC	-0.046, 95% CI [-0.141, 0.044], DeLong $p = 0.331$.
ΔPR AUC	-0.019, 95% CI [-0.090, 0.060].
Δbalanced accuracy	-0.062, 95% CI [-0.125, 0.000].
Δsensitivity	-0.302, 95% CI [-0.412, -0.192].
Δspecificity	0.179, 95% CI [0.106, 0.249].

These selected pairwise results clarify why the main table should be read in terms of tradeoffs rather than rank order alone. Logistic Regression and XGBoost differ in overall ranking performance and in their sensitivity-specificity balance, with Logistic Regression identifying substantially more impaired participants at the cost of lower specificity. Random Forest and GRU-DANN, by contrast, are nearly indistinguishable on ROC AUC alone, yet they operate

in markedly different regions of the decision space: Random Forest is more specific, whereas GRU-DANN is much more sensitive. The most relevant within-family domain-adversarial contrast is DANN versus GRU-DANN. Their ROC AUC difference is not cleanly separated, but GRU-DANN consistently shifts toward higher sensitivity and lower specificity, which explains why it leads balanced accuracy while DANN does not rank first on any of the summary metrics.

4.2 Interpretation of Domain-Adversarial Models

Within this six-model comparison, the domain-adversarial models show different tradeoffs rather than a single clear advantage. GRU-DANN achieves the highest ROC AUC and balanced accuracy, but the absolute ROC AUC remains modest at 0.599. These results should not be interpreted as showing that domain-adversarial learning resolved vehicle heterogeneity. A more cautious reading is that the GRU-DANN configuration performed best on some participant-level metrics in this comparison, while still leaving substantial room for improvement.

Chapter 5

Discussion and Conclusion

5.1 Research Question 1: Discriminative Signal Under LOGO

Research Question 1 asked how much discriminative signal for binary CDR® classification is present under LOGO evaluation, with one individual held out in turn. The main conclusion is that the current driving representation carries only limited signal under this held-out evaluation protocol. Chapter 4 shows modest participant-level performance throughout, so the thesis does not support a strong predictive or clinical claim. That result is still informative. It suggests that weekly aggregated driving summaries retain some behaviorally relevant information, but not enough to separate impaired and unimpaired participants with robust discrimination across the heterogeneous cohort studied here.

The more important interpretation is not which model ranked first on a given metric, but what the weak overall discrimination implies about the representation itself. Chapter 2 argued that extracted telematics features are only useful if they reflect stable behavioral constructs rather than platform, sensing, or contextual quirks. Under LOGO, the current weekly summaries appear to compress substantial variability into a form that is easier to compare but harder to interpret as a durable participant-level signal. In that sense, the results point less to a simple lack of information in naturalistic driving and more to a mismatch between the richness of the underlying behavior and the coarse weekly representation used to summarize it.

5.2 Research Question 2: Domain-Adversarial Modeling

Research Question 2 asked whether treating vehicle-related variation as a domain shift problem changes classification behavior relative to non-adversarial baselines. The results support a narrower answer than the motivating hypothesis. Domain-adversarial modeling did change the error tradeoffs, but Chapter 4 does not show a broad or decisive benefit from DANN under the thesis comparison. DANN does not lead any of the primary Chapter 4 metrics, and its pattern of results suggests limited benefit from adversarial pressure in the present setup rather than a clear demonstration that vehicle-linked nuisance variation was removed.

The more interesting contrast is between flat DANN and GRU-DANN. The Chapter 4 tradeoffs are consistent with the idea that sequence-aware modeling may preserve information that flat weekly summaries blur. Repeated observations contain ordering, persistence, and within-participant fluctuation that are largely flattened once the data are reduced to static weekly aggregates. A model that retains sequence structure can therefore plausibly change detection-oriented behavior even when overall discrimination remains modest. That interpretation should still remain cautious: the current results support saying that sequence-aware modeling may improve some tradeoffs, especially sensitivity-related ones, not that the model recovered a stable temporal impairment signature.

The limited benefit from DANN also has a plausible methodological explanation. Chapter 2 motivated domain adaptation by noting that vehicle heterogeneity can arise through powertrain differences, chassis and braking dynamics, sensing environments, and driver assistance systems. The current thesis does not model that full structure directly. Instead, the adversarial signal is tied to a four-cluster vehicle grouping that is useful as a first approximation but likely too coarse to represent all of the nuisance variation that matters. If the domain labels only partially capture the heterogeneity in the data, then adversarial training cannot be expected to remove it cleanly. More importantly, some vehicle-linked variation may remain entangled with subject-specific driving behavior, so stronger invariance pressure can also discard information that still helps the clinical task.

5.3 Research Question 3: Most Promising Modeling Directions

Research Question 3 asked which combinations of features, demographics, vehicle context, and model design appear most promising for future work on more vehicle-robust driving measures. The evidence points to a selective rather than sweeping answer. The strongest next steps are not simply to search for a new winning classifier, but to improve the representation of behavior, preserve sequence structure where possible, and define vehicle context in a way that more faithfully reflects the sources of nuisance variation present in the cohort.

The sequence result is especially useful in that respect. Even without supporting a strong performance claim, it suggests that ordered longitudinal information may matter because repeated weeks can capture persistence, instability, and change within a participant that a flat summary tends to smooth away. If the target signal is subtle and distributed over time, then preserving temporal structure is a more plausible route forward than relying exclusively on static weekly aggregates. That does not prove that the current GRU-DANN formulation is the correct final model, but it does suggest that future progress is more likely to come from stronger temporal representations than from simply increasing adversarial strength.

The vehicle-context result is similarly interpretive. The four-cluster grouping provides a practical way to acknowledge that vehicle-linked heterogeneity exists, but the limitations discussion already shows why that grouping is incomplete. Residual nuisance variation likely remains in factors such as powertrain behavior, braking and steering response, sensing environment, ADAS availability or activation, and context-dependent interactions between drivers and their vehicles. A more vehicle-robust pipeline will therefore likely require richer context definitions rather than assuming that a coarse cluster label fully captures transportability risk.

The feature-set result should also be stated modestly. The retained 84-feature behavioral subset remains a reasonable frozen representation choice, especially because earlier feature-set work consistently favored removing evidently unhelpful variability families. However, Chapter 4 and the present discussion do not show that this representation is sufficient on its own. The most plausible path forward is therefore cumulative: keep the pruned and interpretable behavioral base, preserve demographic adjustment where appropriate, and pair

those features with models and context definitions that better retain temporal structure and vehicle-related uncertainty.

5.4 Strengths and Limitations

5.4.1 Strengths

One strength of this study is its evaluation discipline. Chapter 3 keeps the distinction between within-family ablation and cross-model comparison explicit, and Chapter 4 keeps participant-level reporting primary. A second strength is that the discussion avoids overstating a single best model when the evidence does not support such a claim. That restraint matters because the main value of the study is not only in what worked, but also in clarifying what still does not generalize cleanly.

5.4.2 Limitations

Several limitations remain. Most importantly, participant-level performance is modest overall, so the current pipeline does not support a strong predictive or clinical claim. The feature construction also relies heavily on weekly aggregation, which may smooth away temporal or contextual patterns that matter for subtle cognitive change. In addition, the nuisance domain is represented only through a four-cluster vehicle-domain grouping, while real vehicle heterogeneity likely reflects a more complex interaction among vehicle characteristics, assistance systems, sensing conditions, and driving context.

The ADAS summaries help make that limitation more concrete, but they do not remove it. Across the 428 VIN-linked vehicles, observed ADAS availability is uneven. When values greater than zero are treated as observed availability, Anti-lock Braking System (46.7%), Electronic Stability Control (44.9%), and Backup Camera (44.6%) are relatively common, while Rear Automatic Emergency Braking (4.0%), Blind Spot Intervention (3.5%), and Lane Centering Assistance (4.2%) are sparse. These values are descriptive only. In the vPIC-based source, 1 means standard, 0.5 means optional, and 0 means not equipped or missing; two fields also contain literal `Not Available` entries.

The K=4 cluster summaries also show descriptive heterogeneity rather than a resolved domain structure. The four vehicle clusters contain 64, 74, 163, and 127 vehicles, and their ADAS profiles differ. However, these fields describe vPIC specification-level availability, not confirmed installed equipment on each study vehicle. For that reason, the ADAS summaries are used only as descriptive context.

More broadly, vehicle characteristics and participant identity are not fully separable in this dataset, so some vehicle-linked effects may remain entangled with subject-specific driving behavior.

5.5 Future Directions

The clearest next step is to improve the behavioral representation rather than assume that stronger adversarial pressure alone will fix the problem. Richer trip-level or route-level representations, and more expressive sequence models that preserve longitudinal structure, are natural priorities because the results suggest that sequence information may matter. A second direction is to refine how vehicle heterogeneity is represented. If the current four-cluster vehicle-domain grouping only partially captures nuisance variation, future domain-adaptation work may need more faithful vehicle-context definitions before its value can be judged fairly. The ADAS descriptive package reinforces that point: future vehicle-context definitions should preserve uncertainty about ambiguous zero-coded fields, keep literal `Not Available` values explicit where they occur, and distinguish cohort-wide prevalence from cluster-wise composition rather than collapsing everything into a binary equipped versus not equipped framing. A third direction concerns evaluation itself. Future improvements should still be checked under LOGO, with GroupShuffleSplit kept only for within-family screening.

5.6 Conclusion

This study asked whether naturalistic driving data can support binary CDR® classification while accounting for vehicle heterogeneity in real-world deployment. The overall answer remains cautious. The results show modest discrimination and clear metric tradeoffs, and they suggest that sequence-based modeling may warrant further study, but they do not

support a deployment-ready clinical claim. They also do not support the stronger conclusion that domain-adversarial learning fully addressed vehicle-linked heterogeneity.

Overall, the results suggest that sequence-based modeling may be more promising than DANN in the present setup, while strong baseline tabular models remain important reference points. The main contribution is therefore an empirical account of what the current pipeline can support, what remains limited, and which next steps appear most plausible for developing more vehicle-robust driving measures.

References

[1] N. Al-Hammadi, M. Abouelyazid, D. C. Brown, P. Lalwani, H. Devos, D. B. Carr, and G. M. Babulal. Integrating Machine Learning and Environmental and Genetic Risk Factors for the Early Detection of Preclinical Alzheimer's Disease. *The Journals of Gerontology, Series B: Psychological Sciences and Social Sciences*, 80(7):gbaf023, June 2025.

[2] F. Al-Hindawi, T. Wu, Y. Wen, P. Serhan, E. Forzani, F. Tsow, and Y. E. Geda. Leveraging naturalistic driving digital biomarkers for early mild cognitive impairment detection: Deep learning strategies. *JMIR Medical Informatics*, 14:e83622, 2026.

[3] G. M. Babulal, A. Johnson, A. M. Fagan, J. C. Morris, and C. M. Roe. Identifying Preclinical Alzheimer's Disease Using Everyday Driving Behavior: Proof of Concept. *Journal of Alzheimer's Disease*, 79(3):1009–1014, Feb. 2021.

[4] J. R. Bacci, S. Karagianni, Z.-S. Alexopoulou, S. Moukaled, C. Tato-Fernández, P. Arunachalam, A. Aslanyan, S. Aye, A. S. L. Rocha, M. Crugel, A. Fawad, A. Sogorb-Esteve, M. Schöll, A. König, and R. W. Paterson. Clinical translation of fluid, imaging, and digital biomarkers for alzheimer's disease. *Alzheimer's Research & Therapy*, 18(1), 2026.

[5] S. Bayat, C. M. Roe, S. Schindler, S. A. Murphy, J. M. Doherty, A. M. Johnson, A. Walker, B. M. Ances, J. C. Morris, and G. M. Babulal. Everyday Driving and Plasma Biomarkers in Alzheimer's Disease: Leveraging Artificial Intelligence to Expand Our Diagnostic Toolkit. *Journal of Alzheimer's Disease*, 92(4):1487–1497, Apr. 2023.

[6] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan. A theory of learning from different domains. *Machine Learning*, 79(1-2):151–175, May 2010.

[7] M. Blake, D. C. Brown, C. Chen, N. Al-Hammadi, R. Casanova, Y. Zhu, and G. M. Babulal. A combined naturalistic driving, clinical, and neurobehavioral data set for investigating aging and dementia. *Scientific Data*, 12(1):1209, July 2025.

[8] Z. Breijyeh and R. Karaman. Comprehensive Review on Alzheimer's Disease: Causes and Treatment. *Molecules*, 25(24):5789, Dec. 2020.

[9] L. Chato and E. E. Regentova. Survey of transfer learning approaches in the machine learning of digital health sensing data. *Journal of Personalized Medicine*, 13(12):1703, Dec. 2023.

[10] K. N. De Winkel and M. Christoph. Rethinking Advanced Driver Assistance System taxonomies: A framework and inventory of real-world safety performance. *Transportation Research Interdisciplinary Perspectives*, 29:101336, Jan. 2025.

[11] Y. Ganin and V. Lempitsky. Unsupervised Domain Adaptation by Backpropagation, Feb. 2015.

[12] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky. Domain-Adversarial Training of Neural Networks, May 2016.

[13] M. Long, Y. Cao, J. Wang, and M. I. Jordan. Learning Transferable Features with Deep Adaptation Networks. *arXiv preprint arXiv:1502.02791*, 2015.

[14] J. C. Morris. The Clinical Dementia Rating (CDR): Current version and scoring rules. *Neurology*, 43(11):2412, Nov. 1993.

[15] S. E. Polk, F. Öhman, J. Hassenstab, A. König, K. V. Papp, M. Schöll, D. Berron, et al. A scoping review of remote and unsupervised digital cognitive assessments in preclinical alzheimer's disease. *npj Digital Medicine*, 8:266, May 2025.

[16] W. Qi, X. Zhu, B. Wang, Y. Shi, C. Dong, S. Shen, J. Li, K. Zhang, Y. He, M. Zhao, S. Yao, Y. Dong, H. Shen, J. Kang, X. Lu, G. Jiang, L. M. M. Boots, H. Fu, L. Pan, H. Chen, Z. Yan, G. Xing, and S. Cao. Alzheimer's disease digital biomarkers multidimensional landscape and ai model scoping review. *npj Digital Medicine*, 8:366, June 2025.

[17] M. Siami, M. Naderpour, and J. Lu. A Mobile Telematics Pattern Recognition Framework for Driving Behavior Extraction. *IEEE Transactions on Intelligent Transportation Systems*, 22(3):1459–1472, Mar. 2021.

[18] T. Varghese, R. Sheelakumari, J. S. James, and P. Mathuranath. A review of neuroimaging biomarkers of Alzheimer's disease. *Neurology Asia*, 18(3):239–248, 2013.

[19] X. Yao, S. C. Calvert, and S. P. Hoogendoorn. Driving heterogeneity identification using machine learning: A review and framework for analysis. *Transportation Research Interdisciplinary Perspectives*, 32:101511, July 2025.

[20] Y. Zheng. Trajectory Data Mining: An Overview. *ACM Transactions on Intelligent Systems and Technology*, 6(3):1–41, May 2015.

Appendix A

Feature Dictionary

This appendix lists the variables used in the thesis feature registry. The full behavioral pool contains 116 variables. The primary behavioral subset retains 84 of those variables and excludes four geographic anchor coordinates and 28 exponentially weighted moving average (EWMA) variability variables. The first four tables list that retained 84-feature behavioral subset, the next two tables list the 32 excluded behavioral variables from the full pool, and the four approved demographic covariates are listed separately at the end of the appendix.

A.1 Behavioral Features in the Primary 84-Feature Subset

Tables A.1–A.4 enumerate the 84 behavioral variables retained from the full 116-variable behavioral pool for the primary analyses.

A.1.1 Speed, acceleration, and event counts

Table A.1: Speed, acceleration, and event counts feature dictionary

Variable	One-line meaning
speed_mean	Average trip-level vehicle speed during the week.
speed_median	Median trip-level vehicle speed during the week.
speed_std	Average within-trip standard deviation of vehicle speed.
speed_min	Average minimum speed observed within retained trips.

Continued on next page

Variable	One-line meaning
speed_max	Average maximum speed observed within retained trips.
speed_percentile_85	Average 85th-percentile speed within retained trips.
hard_braking	Count of hard braking events with deceleration between 8 and 12 mph/s in magnitude.
hardcore_braking	Count of extreme braking events with deceleration greater than 12 mph/s in magnitude.
sudden_acceleration	Count of sudden acceleration events exceeding 8 mph/s while the vehicle is already moving.
braking_count_4	Count of deceleration events with magnitude at least 4 mph/s.
braking_count_5	Count of deceleration events with magnitude at least 5 mph/s.
braking_count_6	Count of deceleration events with magnitude at least 6 mph/s.
braking_count_7	Count of deceleration events with magnitude at least 7 mph/s.
braking_count_8	Count of deceleration events with magnitude at least 8 mph/s.
braking_count_9	Count of deceleration events with magnitude at least 9 mph/s.
braking_count_10	Count of deceleration events with magnitude at least 10 mph/s.
braking_count_11	Count of deceleration events with magnitude at least 11 mph/s.
braking_count_12	Count of deceleration events with magnitude at least 12 mph/s.
accel_count_4	Count of acceleration events with magnitude at least 4 mph/s.
accel_count_5	Count of acceleration events with magnitude at least 5 mph/s.
accel_count_6	Count of acceleration events with magnitude at least 6 mph/s.
accel_count_7	Count of acceleration events with magnitude at least 7 mph/s.
accel_count_8	Count of acceleration events with magnitude at least 8 mph/s.
accel_count_9	Count of acceleration events with magnitude at least 9 mph/s.
accel_count_10	Count of acceleration events with magnitude at least 10 mph/s.
accel_count_11	Count of acceleration events with magnitude at least 11 mph/s.
accel_count_12	Count of acceleration events with magnitude at least 12 mph/s.

A.1.2 Jerk-derived features

Table A.2: Jerk-derived features feature dictionary

Variable	One-line meaning
speed_jerk_bucket_1	Count of absolute speed-jerk values between 5 and 10 mph/s^2.
speed_jerk_bucket_2	Count of absolute speed-jerk values between 10 and 15 mph/s^2.
speed_jerk_bucket_3	Count of absolute speed-jerk values between 15 and 20 mph/s^2.
speed_jerk_bucket_4	Count of absolute speed-jerk values of at least 20 mph/s^2.
speed_jerk_median	Median absolute speed jerk computed from changes in speed over time.
speed_jerk_std	Standard deviation of absolute speed jerk.
rpm_jerk_bucket_1	Count of absolute engine-RPM jerk values between 500 and 1000 RPM/s^2.
rpm_jerk_bucket_2	Count of absolute engine-RPM jerk values between 1000 and 1500 RPM/s^2.
rpm_jerk_bucket_3	Count of absolute engine-RPM jerk values of at least 1500 RPM/s^2.
rpm_jerk_median	Median absolute engine-RPM jerk.
rpm_jerk_std	Standard deviation of absolute engine-RPM jerk.
accel_jerk_bucket_1	Count of longitudinal accelerometer jerk values between 0 and 0.025 g/s.
accel_jerk_bucket_2	Count of longitudinal accelerometer jerk values between 0.025 and 0.05 g/s.
accel_jerk_bucket_3	Count of longitudinal accelerometer jerk values between 0.05 and 0.1 g/s.
accel_jerk_bucket_4	Count of longitudinal accelerometer jerk values of at least 0.1 g/s.
accel_jerk_median	Median longitudinal accelerometer jerk computed over 0.5-second intervals.
accel_jerk_std	Standard deviation of longitudinal accelerometer jerk over 0.5-second intervals.

A.1.3 Route, turning, and stop behavior

Table A.3: Route, turning, and stop behavior feature dictionary

Variable	One-line meaning
corneringCount	Count of cornering events carried through from the study metrics dataset.
displacement_ratio	Straight-line start-to-end distance divided by total trip distance.
distance_miles	Total estimated miles traveled during the week.
gps_turn_count_30	Count of heading changes of at least 30 degrees.
gps_turn_count_60	Count of heading changes of at least 60 degrees.
gps_turn_count_90	Count of heading changes of at least 90 degrees.
heading_change_mean	Average absolute heading change between consecutive GPS points.
heading_change_std	Standard deviation of absolute heading change between consecutive GPS points.
heading_change_total	Total absolute heading change accumulated across the trip path.
heading_entropy	Shannon entropy of trip headings, reflecting directional diversity.
route_complexity_index	Count of heading changes of at least 30 degrees per mile traveled.
mid_trip_stop_count	Number of mid-trip stops lasting at least 30 seconds, excluding the first and last 10% of each trip.
mid_trip_stop_fraction	Fraction of trimmed mid-trip time spent stopped.
mid_trip_stop_mean_duration	Average duration of qualifying mid-trip stops.
mid_trip_stop_total_duration	Total duration of qualifying mid-trip stops.

A.1.4 Driving context and data quality

Table A.4: Driving context and data quality feature dictionary

Variable	One-line meaning
duration_seconds	Total driving time during the week in seconds.
duration_minutes	Total driving time during the week in minutes.
idleTime	Total idling time recorded in the study metrics dataset.
stopTime	Total stop time recorded in the study metrics dataset.
overspeedingCount	Count of overspeeding episodes recorded in the study metrics dataset.
overspeedingDuration	Total overspeeding duration recorded in the study metrics dataset.
missing_accel_pct	Percentage of telemetry records missing accelerometer data.
missing_gps_pct	Percentage of telemetry records missing GPS data.
missing_gyro_pct	Percentage of telemetry records missing gyroscope data.
missing_rpm_pct	Percentage of telemetry records missing engine-RPM data.
missing_speed_pct	Percentage of telemetry records missing speed data.
n_trips	Number of retained trips contributing to the weekly observation.
num_records	Total number of retained telemetry records across trips in the week.
time_gap_count	Count of inter-record gaps greater than 1.5 seconds.
tripCountUpTo1mile	Count of trips shorter than 1 mile.
tripCount1UpTo5miles	Count of trips between 1 and 5 miles.
tripCount5UpTo10miles	Count of trips between 5 and 10 miles.
tripCount10UpTo20miles	Count of trips between 10 and 20 miles.
`tripCount20MilesOr Greater`	Count of trips of at least 20 miles.
tripDawn	Count of trips classified as dawn trips.
tripDaytime	Count of trips classified as daytime trips.
tripDusk	Count of trips classified as dusk trips.
tripNighttime	Count of trips classified as nighttime trips.
tripStartsDuringDay	Weekly flag indicating whether any retained trip started during daytime.
tripEndsDuringDay	Weekly flag indicating whether any retained trip ended during daytime.

A.2 Behavioral Features Excluded from the Primary 84-Feature Subset

The next two tables list the 32 behavioral variables that remain part of the full 116-variable pool but are excluded from the primary 84-feature subset.

A.2.1 GPS anchor variables

Table A.5: GPS anchor variables feature dictionary

Variable	One-line meaning
first_lat	Latitude of the earliest valid trip start in the week.
first_lon	Longitude of the earliest valid trip start in the week.
last_lat	Latitude of the latest valid trip end in the week.
last_lon	Longitude of the latest valid trip end in the week.

A.2.2 EWMA variability variables

Table A.6: EWMA variability variables feature dictionary

Variable	One-line meaning
ewma_sd_d_accel_count_8	EWMA standard deviation of week-to-week change in count of acceleration events with magnitude at least 8 mph/s.
ewma_sd_d_braking_count_8	EWMA standard deviation of week-to-week change in count of deceleration events with magnitude at least 8 mph/s.
ewma_sd_d_displacement_ratio	EWMA standard deviation of week-to-week change in straight-line start-to-end distance divided by total trip distance.
ewma_sd_d_distance_miles	EWMA standard deviation of week-to-week change in total estimated miles traveled during the week.
ewma_sd_d_duration_seconds	EWMA standard deviation of week-to-week change in total driving time during the week in seconds.

Continued on next page

Variable	One-line meaning
ewma_sd_d_hard_braking	EWMA standard deviation of week-to-week change in count of hard braking events with deceleration between 8 and 12 mph/s in magnitude.
ewma_sd_d_hardcore_braking	EWMA standard deviation of week-to-week change in count of extreme braking events with deceleration greater than 12 mph/s in magnitude.
ewma_sd_d_heading_entropy	EWMA standard deviation of week-to-week change in shannon entropy of trip headings, reflecting directional diversity.
ewma_sd_d_speed_jerk_median	EWMA standard deviation of week-to-week change in median absolute speed jerk computed from changes in speed over time.
ewma_sd_d_speed_jerk_std	EWMA standard deviation of week-to-week change in standard deviation of absolute speed jerk.
ewma_sd_d_speed_mean	EWMA standard deviation of week-to-week change in average trip-level vehicle speed during the week.
ewma_sd_d_speed_percentile_85	EWMA standard deviation of week-to-week change in average 85th-percentile speed within retained trips.
ewma_sd_d_speed_std	EWMA standard deviation of week-to-week change in average within-trip standard deviation of vehicle speed.
ewma_sd_d_sudden_acceleration	EWMA standard deviation of week-to-week change in count of sudden acceleration events exceeding 8 mph/s while the vehicle is already moving.
ewma_var_d_accel_count_8	EWMA variance of week-to-week change in count of acceleration events with magnitude at least 8 mph/s.
ewma_var_d_braking_count_8	EWMA variance of week-to-week change in count of deceleration events with magnitude at least 8 mph/s.
ewma_var_d_displacement_ratio	EWMA variance of week-to-week change in straight-line start-to-end distance divided by total trip distance.
ewma_var_d_distance_miles	EWMA variance of week-to-week change in total estimated miles traveled during the week.
ewma_var_d_duration_seconds	EWMA variance of week-to-week change in total driving time during the week in seconds.

Continued on next page

Variable	One-line meaning
ewma_var_d_hard_braking	EWMA variance of week-to-week change in count of hard braking events with deceleration between 8 and 12 mph/s in magnitude.
ewma_var_d_hardcore_braking	EWMA variance of week-to-week change in count of extreme braking events with deceleration greater than 12 mph/s in magnitude.
ewma_var_d_heading_entropy	EWMA variance of week-to-week change in shannon entropy of trip headings, reflecting directional diversity.
ewma_var_d_speed_jerk_median	EWMA variance of week-to-week change in median absolute speed jerk computed from changes in speed over time.
ewma_var_d_speed_jerk_std	EWMA variance of week-to-week change in standard deviation of absolute speed jerk.
ewma_var_d_speed_mean	EWMA variance of week-to-week change in average trip-level vehicle speed during the week.
ewma_var_d_speed_percentile_85	EWMA variance of week-to-week change in average 85th-percentile speed within retained trips.
ewma_var_d_speed_std	EWMA variance of week-to-week change in average within-trip standard deviation of vehicle speed.
ewma_var_d_sudden_acceleration	EWMA variance of week-to-week change in count of sudden acceleration events exceeding 8 mph/s while the vehicle is already moving.

A.3 Demographic Covariates

These four variables are not part of the 116-variable behavioral pool, but they are the approved demographic covariates for the thesis demographic-control analyses. This appendix documents the encoded demographic covariates used in the thesis pipeline, but it does not assign human-readable semantics to raw demographic study codes because the repository does not include the authoritative production codebook for those source-data codes.

Table A.7: Demographic covariate dictionary

Variable	One-line meaning
age_at_cdr	Participant age at the time of the closest retained CDR®assessment.
educ	Years of education.
gender_encoded	Encoded gender covariate used in the demographic control set.
race_encoded	Encoded race covariate used in the demographic control set.

A.4 Vehicle-Specification Features Used for Vehicle-Domain Labels

This appendix also lists the non-behavioral vehicle-specification features used to derive the four-cluster vehicle-domain grouping referenced in Chapter 3 and Chapter 5. These variables are not part of the 116-variable behavioral pool. Instead, they are used only to construct the nuisance-domain labels for the domain-adversarial models and to provide descriptive vehicle-context information.

The vehicle-domain grouping is built from 24 vehicle-specification features: 20 ADAS availability fields together with `DriveType`, `BodyType`, `ModelYear`, and `Cylinders`. In the NHTSA vPIC-derived source, ADAS coding is specification-level rather than inspection-level: `1` means standard, `0.5` means optional, and `0` means not equipped or missing. In addition, `PedestrianAutomaticEmergencyBraking` and `ParkingAssist` can contain literal `Not Available` values. These fields are therefore used here only to define vehicle-context groupings and to provide descriptive context, not as confirmed equipment records for each study vehicle.

Table A.8: Vehicle-specification features used to derive vehicle-domain labels

Variable	One-line meaning
ADAS availability fields	
CrashImminent	Crash Imminent Braking availability in the vPIC-derived source.
RearAutomaticEmergencyBraking	Rear Automatic Emergency Braking availability in the vPIC-derived source.
PedestrianAutomaticEmergencyBraking	Pedestrian Automatic Emergency Braking availability in the vPIC-derived source.
AntiLockBrakingSystem	Anti-lock Braking System availability in the vPIC-derived source.
ElectronicStabilityControl	Electronic Stability Control availability in the vPIC-derived source.
TractionControl	Traction Control availability in the vPIC-derived source.
ForwardCollisionWarning	Forward Collision Warning availability in the vPIC-derived source.
BlindSpotWarning	Blind Spot Warning availability in the vPIC-derived source.
BlindSpotIntervention	Blind Spot Intervention availability in the vPIC-derived source.
LaneKeepingAssistance	Lane Keeping Assistance availability in the vPIC-derived source.
LaneCenteringAssistance	Lane Centering Assistance availability in the vPIC-derived source.
RearCrossTrafficAlert	Rear Cross Traffic Alert availability in the vPIC-derived source.
AdaptiveCruiseControl	Adaptive Cruise Control availability in the vPIC-derived source.
DynamicBrakeSupport	Dynamic Brake Support availability in the vPIC-derived source.
BackupCamera	Backup Camera availability in the vPIC-derived source.

Continued on next page

Variable	One-line meaning
ParkingAssist	Parking Assist availability in the vPIC-derived source.
DaytimeRunningLight	Daytime Running Light availability in the vPIC-derived source.
AdaptiveDrivingBeam	Adaptive Driving Beam availability in the vPIC-derived source.
SemiautomaticHeadlampBeamSwitching	Semiautomatic Headlamp Beam Switching availability in the vPIC-derived source.
KeylessIgnition	Keyless Ignition availability in the vPIC-derived source.
Other vehicle-specification fields	
DriveType	Drivetrain category used in vehicle clustering, such as FWD, RWD, AWD, or 4WD.
BodyType	Vehicle body classification used in vehicle clustering, such as sedan, SUV, truck, coupe, or wagon.
ModelYear	Vehicle model year used as a numeric clustering feature.
Cylinders	Engine cylinder count used as a numeric clustering feature.

Appendix B

Pairwise Participant-Level Results

Tables B.1 and B.2 provide the full participant-level pairwise comparison surface used to support the Chapter 4 interpretation. Each row reports model A minus model B on the same aligned held-out participants.

B.1 Pairwise ROC AUC Comparison

Table B.1 gives the ROC AUC differences with paired bootstrap confidence intervals and paired DeLong p-values.

Table B.1: Full pairwise ROC AUC comparison table for the selected six-model Chapter 4 comparison. Δ is reported as model A minus model B, and $\uparrow$ indicates higher is better.

Model pair	ΔAUC $\uparrow$	95% paired CI	DeLong p
Logistic Regression vs. XGBoost	0.084	[0.008, 0.152]	0.027
Logistic Regression vs. Random Forest	-0.010	[-0.076, 0.055]	0.782
Logistic Regression vs. MLP	0.013	[-0.062, 0.082]	0.736
Logistic Regression vs. DANN	0.032	[-0.044, 0.109]	0.409
Logistic Regression vs. GRU-DANN	-0.014	[-0.095, 0.064]	0.727
XGBoost vs. Random Forest	-0.093	[-0.147, -0.043]	0.001
XGBoost vs. MLP	-0.071	[-0.142, 0.005]	0.060
XGBoost vs. DANN	-0.052	[-0.122, 0.020]	0.128
XGBoost vs. GRU-DANN	-0.097	[-0.183, -0.004]	0.033
Random Forest vs. MLP	0.022	[-0.049, 0.095]	0.549
Random Forest vs. DANN	0.041	[-0.033, 0.120]	0.279
Random Forest vs. GRU-DANN	-0.004	[-0.085, 0.078]	0.920

Continued on next page

Model pair	ΔAUC ↑	95% paired CI	DeLong p
MLP vs. DANN	0.019	[-0.029, 0.066]	0.432
MLP vs. GRU-DANN	-0.027	[-0.114, 0.058]	0.563
DANN vs. GRU-DANN	-0.046	[-0.141, 0.044]	0.331

B.2 Pairwise Bootstrap Comparison

Table B.2 gives the paired bootstrap differences for PR AUC, balanced accuracy, sensitivity, and specificity. Positive Δ values favor model A for every metric shown; ↑ indicates higher is better.

Table B.2: Full paired bootstrap comparison table for PR AUC, balanced accuracy, sensitivity, and specificity in the selected six-model Chapter 4 comparison. Δ is reported as model A minus model B, and ↑ indicates higher is better.

Model pair	Metric	Δ ↑	95% paired CI
Logistic Regression vs. XGBoost	PR AUC	0.053	[-0.017, 0.119]
	Bal. Acc.	0.070	[-0.003, 0.135]
	Sensitivity	0.314	[0.181, 0.423]
	Specificity	-0.174	[-0.245, -0.101]
Logistic Regression vs. Random Forest	PR AUC	-0.011	[-0.085, 0.054]
	Bal. Acc.	0.041	[-0.028, 0.104]
	Sensitivity	0.326	[0.208, 0.432]
	Specificity	-0.243	[-0.308, -0.177]
Logistic Regression vs. MLP	PR AUC	0.007	[-0.067, 0.077]
	Bal. Acc.	0.049	[-0.018, 0.112]
	Sensitivity	0.337	[0.217, 0.444]
	Specificity	-0.239	[-0.305, -0.170]
Logistic Regression vs. DANN	PR AUC	0.017	[-0.057, 0.082]
	Bal. Acc.	0.054	[-0.012, 0.115]
	Sensitivity	0.314	[0.200, 0.421]
	Specificity	-0.206	[-0.277, -0.133]

Continued on next page

Model pair	Metric	Δ ↑	95% paired CI
Logistic Regression vs. GRU-DANN	PR AUC	-0.002	[-0.083, 0.069]
	Bal. Acc.	-0.008	[-0.092, 0.065]
	Sensitivity	0.012	[-0.133, 0.148]
	Specificity	-0.028	[-0.098, 0.052]
XGBoost vs. Random Forest	PR AUC	-0.063	[-0.114, -0.022]
	Bal. Acc.	-0.029	[-0.072, 0.013]
	Sensitivity	0.012	[-0.069, 0.085]
	Specificity	-0.069	[-0.112, -0.028]
XGBoost vs. MLP	PR AUC	-0.045	[-0.115, 0.023]
	Bal. Acc.	-0.020	[-0.073, 0.038]
	Sensitivity	0.023	[-0.071, 0.122]
	Specificity	-0.064	[-0.122, -0.009]
XGBoost vs. DANN	PR AUC	-0.035	[-0.102, 0.020]
	Bal. Acc.	-0.016	[-0.066, 0.039]
	Sensitivity	0.000	[-0.092, 0.102]
	Specificity	-0.032	[-0.084, 0.023]
XGBoost vs. GRU-DANN	PR AUC	-0.054	[-0.132, 0.028]
	Bal. Acc.	-0.078	[-0.148, -0.001]
	Sensitivity	-0.302	[-0.431, -0.169]
	Specificity	0.147	[0.073, 0.219]
Random Forest vs. MLP	PR AUC	0.018	[-0.052, 0.087]
	Bal. Acc.	0.008	[-0.042, 0.060]
	Sensitivity	0.012	[-0.083, 0.103]
	Specificity	0.005	[-0.044, 0.056]
Random Forest vs. DANN	PR AUC	0.028	[-0.039, 0.095]
	Bal. Acc.	0.013	[-0.038, 0.066]
	Sensitivity	-0.012	[-0.105, 0.084]
	Specificity	0.037	[-0.010, 0.091]
Random Forest vs. GRU-DANN	PR AUC	0.009	[-0.070, 0.091]
	Bal. Acc.	-0.049	[-0.118, 0.022]
	Sensitivity	-0.314	[-0.435, -0.188]
	Specificity	0.216	[0.152, 0.284]

Continued on next page

Model pair	Metric	$\Delta \uparrow$	95% paired CI
MLP vs. DANN	PR AUC	0.010	[-0.033, 0.051]
	Bal. Acc.	0.004	[-0.025, 0.033]
	Sensitivity	-0.023	[-0.074, 0.021]
	Specificity	0.032	[-0.005, 0.073]
MLP vs. GRU-DANN	PR AUC	-0.009	[-0.083, 0.076]
	Bal. Acc.	-0.057	[-0.123, 0.011]
	Sensitivity	-0.326	[-0.447, -0.202]
	Specificity	0.211	[0.145, 0.276]
DANN vs. GRU-DANN	PR AUC	-0.019	[-0.090, 0.060]
	Bal. Acc.	-0.062	[-0.125, 0.000]
	Sensitivity	-0.302	[-0.412, -0.192]
	Specificity	0.179	[0.106, 0.249]