Int. J. Human–Computer Studies 213 (2026) 103827



Contents lists available at ScienceDirect

International Journal of Human - Computer Studies

journal homepage: www.elsevier.com/locate/ijhcs



Gaze dynamics reveal age-related physiological patterns across driving events

Gregor Strle[a,b,*], Kristina Stojmenova Pečečnik[a], Jaka Sodnik[a]

[a] Faculty of Electrical Engineering, University of Ljubljana, Tržaška 25, Ljubljana, 1000, Slovenia
[b] Scientific Research Centre ZRC SAZU, Novi trg 2, Ljubljana, 1000, Slovenia

HIGHLIGHTS

- Gaze features achieve 82.4% accuracy in discriminating four driving events.
- Older drivers show elevated pupil motif amplitudes during constrained gaze.
- Horizontal gaze deviation variability is the strongest event discriminator.
- Time-series motifs reveal age-specific physiological response patterns.
- Convergent methods confirm robust gaze-based multi-event discrimination.

ARTICLE INFO

Keywords:
Gaze dynamics
Physiological patterns
Aging
Driver behavior
Driver monitoring systems
Machine learning
Time-series analysis

ABSTRACT

Discriminating between multiple driving events using physiological signals remains challenging. This study examined whether gaze dynamics and pupil responses could distinguish multiple driving events and reveal age-related physiological patterns. Twenty-seven participants (12 young, 15 old) completed a simulated driving scenario featuring nine events while physiological signals were recorded. Statistical analyses revealed significant event-specific modulation of gaze dispersion metrics ($\eta^2 = 0.571$–0.582, $p < 0.001$). Gradient boosted trees achieved 82.4% accuracy (95% CI [74.3%, 89.8%]) in classifying four events (Bicycle, DeerAlert, LowGas, StopAtGasStation) using 7 gaze-based features, with SHAP analysis identifying horizontal deviation standard deviation and median as primary discriminators. Time-series motif discovery revealed consistent gaze motif amplitudes (3.0–3.4 z-scores) across events, while pupil motifs showed selective enhancement during monitoring tasks (3.1–3.8 z-scores). Age-stratified analyses uncovered distinct physiological patterns: younger drivers exhibited greater gaze variability with broader scanning patterns, whereas older drivers demonstrated elevated pupil motif amplitudes (particularly during LowGas: 3.88 vs. 3.18 z-scores) alongside more constrained visual exploration. The convergence of statistical, machine learning, and motif discovery approaches establishes gaze dynamics as sufficient for multiclass event discrimination in simulated driving. These findings demonstrate that gaze-based features alone can reliably distinguish between specific driving events, while motif analysis reveals temporal dynamics and age-related patterns that warrant further investigation in larger samples, providing a foundation for interpretable driver monitoring systems.

1. Introduction

Modern vehicles increasingly rely on driver monitoring systems to enhance road safety, yet these systems remain limited in their ability to understand the nuanced physiological responses that characterize different driving scenarios. Current technologies primarily focus on detecting binary states – alert versus drowsy, attentive versus distracted – using cameras, steering sensors, or physiological measures. While these binary classifications achieve high accuracy, real-world driving demands recognition of specific events that require distinct driver responses: a pedestrian crossing requires different visual and motor patterns than monitoring a fuel warning or responding to sudden wildlife hazards. This capability gap constrains the development of context-aware safety systems that could provide targeted, event-specific interventions.

* Corresponding author at: Faculty of Electrical Engineering, University of Ljubljana, Tržaška 25, Ljubljana, 1000, Slovenia.
Email addresses: gregor.strle@fe.uni-lj.si (G. Strle), Kristina.Stojmenova@fe.uni-lj.si (K.S. Pečečnik), Jaka.Sodnik@fe.uni-lj.si (J. Sodnik).

https://doi.org/10.1016/j.ijhcs.2026.103827
Received 17 September 2025; Received in revised form 16 April 2026; Accepted 28 April 2026
Available online 5 May 2026
1071-5819/© 2026 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

Abbreviations

CNN Convolutional Neural Network
CRUNCH Compensation-Related Utilization of Neural Circuits Hypothesis
DBA DTW Barycenter Averaging
DTW Dynamic Time Warping
EEG Electroencephalography
EEGNet Convolutional Neural Network tailored for EEG classification
FFT Fast Fourier Transform
HUD Head-Up Display
IoU Intersection-over-Union
IVIS In-Vehicle Infotainment System
LOPO-CV Leave-One-Participant-Out Cross-Validation
LSTM Long Short-Term Memory
MMPNet Multimodal Physiological Pattern Network
NASA-TLX NASA Task Load Index
PRT Probe Reaction Time
PR-AUC Precision–Recall – Area Under the Curve
ROC-AUC Receiver Operating Characteristic – Area Under the Curve
SHAP SHapley Additive exPlanations
SVM Support Vector Machine
XGB eXtreme Gradient Boosting

Physiological signals, particularly ocular measures, provide a continuous, non-invasive window into driver cognitive state. Eye movements and pupil responses reflect moment-to-moment changes in attention allocation, cognitive load, and arousal levels, with theoretical foundations establishing their link to underlying cognitive processes (Yarbus, 1967; Kahneman, 1973). The automotive industry has embraced eye-tracking technology, with systems now deployed in production vehicles for drowsiness and attention monitoring. Recent meta-analyses confirm that binary classification using ocular features achieves 85–95% accuracy in controlled environments (Koay et al., 2022; Michelaraki et al., 2023). However, the potential for these same physiological signals to discriminate between multiple specific driving events remains largely unexplored. The few attempts at event-specific detection have been limited to distinguishing between two hazard types or predicting single events like lane changes, leaving it uncertain whether physiological patterns can reliably identify diverse driving scenarios (Doshi and Trivedi, 2009).

Age-related changes in driving behavior add complexity to physiological monitoring approaches. Older drivers consistently demonstrate altered visual scanning strategies, with 28–35% smaller horizontal scan paths and fewer peripheral glances compared to younger drivers (Maltz and Shinar, 1999; Underwood, 2007). These behavioral changes are typically framed as deficits requiring compensation through vehicle automation or warning systems. However, whether older drivers develop alternative physiological strategies to maintain situational awareness despite constrained visual exploration remains unexplored. Recent evidence suggests older adults may enhance certain physiological responses under high cognitive load Kim et al. (2023), but systematic investigation in driving contexts is lacking. Moreover, traditional analysis approaches using aggregated features may miss temporal dynamics that could reveal these compensatory mechanisms.

This study addresses these gaps by examining whether gaze dynamics and pupil responses can discriminate between multiple driving events while revealing potential age-related physiological strategies. Using a controlled simulator environment with standardized event sequences, we employed converging analytical approaches – statistical analyses to establish signal reliability, machine learning classification to quantify discrimination performance, and time-series motif discovery to identify recurring temporal patterns. This multi-method framework enables both validation of event-specific physiological signatures and exploration of how these patterns vary with age.

Specifically, this study addresses three research questions: (1) Can physiological signals, particularly gaze dynamics and pupil responses, discriminate between multiple distinct driving events beyond binary classification? (2) Do older and younger drivers exhibit different physiological response patterns during these events? (3) Can temporal pattern analysis reveal structured behavioral strategies that complement traditional aggregate measures? We hypothesize that gaze-based features will provide sufficient information for multi-event classification, and that age groups will show distinct but potentially complementary physiological strategies.

The following sections present the theoretical and empirical foundations for physiological monitoring in driving (Section 2), describe the experimental methodology and analytical approaches (Section 3), report findings across the three analytical frameworks (Section 4), and discuss implications for driver monitoring system design (Section 5).

2. Related work

The theoretical foundation for using eye movements and pupil responses as cognitive indicators originates from Yarbus's seminal work Yarbus (1967), which demonstrated that scanpaths systematically vary with task demands and goals. This established the principle that visual behavior reflects underlying cognitive processes. Building on this foundation, pupillometry research by Hess and Polt (1964) quantified the relationship between mental effort and pupil size, finding proportional increases with arithmetic problem difficulty in 22 participants. Kahneman and Beatty (1966) extended these findings to memory tasks, documenting systematic pupil enlargement with increasing digit span loads in 12 participants. These empirical observations were synthesized in Kahneman's framework (Kahneman, 1973), establishing pupil dilation as a reliable physiological index of cognitive load and attentional resource allocation.

2.1. Physiological indicators in driving contexts

The application of pupillometry to driving has consistently validated pupil diameter as a cognitive load indicator. For example, Čegovnik et al. (2018) demonstrated this relationship using affordable eye trackers with 24 participants (mean age 27.3 years) performing n-back tasks during simulated driving. Pupil diameter increased significantly from baseline (3.2 mm) to high cognitive load conditions (4.1 mm, $p < 0.001$). Nilsson et al. (2018) conducted a larger study with 70 drivers divided by age (35 younger: 20–30 years, 35 older: 65–75 years). Their results showed stepwise pupil enlargement corresponding to task difficulty: baseline driving (3.5 mm), 1-back task (3.9 mm), and 2-back task (4.3 mm), with pupil size emerging as the most reliable indicator among multiple physiological measures including EEG, heart rate, and skin conductance.

Aggregating across multiple on-road and simulator datasets (total $N \approx 108$), Mehler et al. (2012) showed that physiological measures – most notably pupil diameter – scale with graded working-memory load during driving across environments and age strata. Earlier dual-task experiments comparing 18 younger (19–23 years) and 15 older (51–66 years) drivers reported age-differentiated autonomic responses: younger participants typically exhibited heart-rate acceleration with added load, whereas older drivers as a group showed attenuated or even deceleratory responses (consistent with an orienting response) under the same conditions (Mehler et al., 2008). Related incremental-load work in young drivers established that arousal measures (heart rate, electrodermal

activity, pupil dilation) rise systematically with cognitive demand even when initial driving performance is maintained (Mehler et al., 2009).

Beyond cognitive load assessment, pupil diameter has been validated for detecting the impact of in-vehicle technologies. Wei et al. (2023) tested 30 participants in a simulated urban driving environment with varying levels of auditory memory tasks delivered through in-vehicle infotainment systems. They found that task difficulty significantly increased both pupil diameter and subjective workload scores on NASA-TLX, confirming that pupillometry can sensitively capture cognitive demands induced by secondary IVIS tasks.

Complementing pupillary measures, gaze dispersion patterns have emerged as equally important indicators of cognitive state. Recarte and Nunes (2003) conducted on-road research with 12 experienced drivers (ages 25–40) performing mental arithmetic while driving. Despite maintaining their eyes on the road, participants exhibited 23% reduction in horizontal scan variability and 42% fewer mirror checks during cognitive tasks. This 'tunnel vision' phenomenon has proven robust across driving contexts. Zhang et al. (2025) explored situational trust and visual behavior in advanced driver assistance systems (ADAS), finding that trustful drivers spent less time monitoring the driving environment compared to distrustful drivers, particularly in complex scenarios such as lane addition/reduction. Their on-road study with 34 experienced drivers demonstrated that higher trust levels were associated with reduced environmental monitoring and increased focus on the human-machine interface.

Wang et al. (2014) investigated gaze concentration sensitivity during real-world driving under cognitive load in an instrumented-vehicle study with 35 older licensed drivers (ages 60–75). Participants completed auditory n-back (1-back, 2-back) and visuospatial tasks while driving. Comparing multiple quantification methods – standard deviation of horizontal/vertical gaze position, percent road center (PRC) with various window radii, and fixation-based versus sample-based metrics – they found horizontal gaze position standard deviation was the most sensitive index, showing the largest effect sizes particularly for 2-back versus baseline. Effects were stronger for auditory 2-back than visuospatial tasks, and results remained robust whether saccades were retained or removed, supporting the utility of dispersion-style metrics for detecting workload-related attentional narrowing during naturalistic driving.

A recent meta-analysis of 18 driving studies confirmed that both pupil dilation and ocular metrics show medium-to-large associations with cognitive load manipulations in driving tasks, with pupil size and heart rate providing the strongest associations ($r > 0.5$), while blink rate and respiration showed weaker, context-dependent effects (Wang et al., 2024). Beyond mean pupil diameter, composite temporal metrics capture pupil oscillation dynamics. Duchowski et al. (2020) developed the Low/High Index of Pupillary Activity (LHIPA), which quantifies cognitive load through the ratio of low-frequency to high-frequency pupil oscillations. LHIPA demonstrated superior discrimination of task difficulty compared to traditional mean diameter measures across fixed-gaze counting, n-back, and eye-typing tasks, highlighting the value of temporal pupil dynamics beyond aggregate statistics.

2.2. Age-related adaptations in driving

The broader cognitive and sensory context for age differences in driving emerges from comprehensive analyses of aging and performance. Anstey et al. (2005) integrate evidence across perceptual (e.g., contrast sensitivity, visual fields), cognitive (e.g., processing speed, selective/divided attention, executive function), and physical factors to explain why older adults experience difficulty under higher demand or complexity. Aggregating laboratory, simulator, and on-road studies, they present a multifactorial account where constraints in visual attention and executive control combine with sensory decline to narrow the effective field of view and slow information uptake, setting up conditions for delayed or conservative gaze strategies in complex scenes. Similarly, Charness et al. examine aging and human performance from a human-factors perspective, documenting how age-related changes in attention allocation, working memory, and psychomotor speed affect complex, time-pressured tasks. Their analysis of measurement approaches (from glance metrics to physiological indices) and design implications (e.g., environmental support, interface pacing) establishes a framework in which differences in temporal organization of gaze and arousal with age are expected rather than anomalous (Charness, 2008).

2.2.1. Behavioral changes in visual search strategies

Age-related changes in driving behavior have been extensively documented, with research primarily emphasizing decline in visual search strategies. Maltz and Shinar (1999) conducted a comparative study with 20 older drivers (ages 65–75) and 20 younger drivers (ages 25–35) navigating urban intersections. Their eye-tracking analysis revealed that older drivers made 35% fewer glances at mirrors and side windows, spending significantly more time fixating on the road ahead. This reduction in peripheral monitoring was accompanied by longer fixation durations on each area of interest, suggesting a more serial rather than parallel visual processing strategy.

Underwood (2007) examined visual scanning patterns in 30 drivers across three age groups during navigation of complex traffic scenarios. They reported 28% smaller horizontal scan paths in older drivers, with reduced variance in gaze positions particularly evident during high-demand situations such as merging and overtaking. From a cognitive perspective, Underwood (2007) reviewed how visual attention reorganizes with experience (novice → advanced), emphasizing anticipatory scanning and hazard anticipation – constructs subsequently evaluated in age-comparative studies. While these studies document substantial reductions in visual search efficiency, recent evidence reveals a more nuanced picture.

Recent simulator evidence continues to probe the interaction of age, visual cognition, and intersection behavior. Guo et al. (2025) compared 15 younger/middle-aged drivers (24–47 years) and 15 older drivers (60+) across critical intersection scenarios with varied traffic complexity, collecting eye-movement and driving-performance measures. They reported that age significantly affected visual perception and control – older drivers exhibited more difficulty in information processing but also compensatory strategies such as slower approach speeds – and quantified links between visual-cognition indices and operational performance.

In a driving simulator study that manipulated roadway complexity, Cantin et al. (2009) contrasted younger and older licensed drivers using concurrent probe reaction time (PRT) and driving performance indices (speed regulation, lateral control, braking). Ten younger men (mean ≈ 24 years) and ten older men (mean ≈ 68 years) drove a 26.4 km route with varying curves, speed postings, traffic, and car-following segments; PRTs were elicited during driving and vehicle-control measures were logged. Older adults incurred disproportionately larger workload costs and performance decrements as complexity increased (e.g., longer PRTs under high demand, more frequent hard-braking events), consistent with age-linked changes in how scanning and control are temporally organized under pressure (Cantin et al., 2009).

In a PRT/eye-movement study focused on intersections, Yamani et al. (2016) analyzed middle-aged versus older drivers and found that older participants executed secondary glances at intersections only about half as often as the middle-aged group and showed more centralized gaze when managing peripheral visual search while maintaining lane position. Their sample comprised 12 middle-aged (5 female; mean 58.3 years, SD 4.75, range 50–65; mean far acuity 20/21.0; mean maximum head rotation 59.3° ± 12.0°) and 12 older drivers (9 female; mean 75.7 years, SD 4.11, range 70–85; mean far acuity 20/26.7; maximum head rotation 55.3° ± 10.1°); all held valid licenses and reported substantial annual mileage (middle-aged: ≈21,500 km; older: ≈15,700 km) (Yamani et al., 2016). The authors discuss these results in light of the 'decoupling hypothesis', whereby older drivers may couple multi-limb responses (steering/acceleration) with head/eye movements

during turning, contributing to neglected secondary glances at critical times (Yamani et al., 2016).

2.2.2. Compensatory mechanisms and neural adaptations

Beyond documenting behavioral changes, emerging evidence suggests older adults may employ compensatory mechanisms at multiple levels. At the neural level, Getzmann et al. (2018) investigated whether older drivers recruit additional neural resources to maintain performance during lane-keeping. Sixteen younger (8 female; mean 24.1 years, range 20–30) and sixteen older (8 female; mean 63.3 years, range 55–69) drivers participated; one older participant was excluded due to EEG artifacts. Despite large differences in years licensed (young: 6.7; older: 43.7), mean annual mileage did not differ materially (young ≈12,207 km; older ≈12,455 km) (Getzmann et al., 2018). Older adults exhibited significantly higher frontal theta power (4–7 Hz) under increasing workload, and within this group, higher theta activity correlated with more stable steering performance ($\tau \approx -0.5$, $p < 0.03$). This pattern, absent in younger adults, suggests that steadier control coincided with increased mental effort, consistent with compensatory recruitment of frontal control resources.

Similar compensatory patterns emerge at the oculomotor level. Zargari Marandi et al. (2018) examined how eye movement characteristics reflect fatigue development in young (n = 17) and older (n = 17) adults during a two-hour sustained attention task. Older adults had longer fixation durations ($F_{1,32} = 14.8$, $p = 0.001$) and shorter saccade durations, reflecting a slower and more deliberate scanning strategy. Despite these baseline differences, both groups displayed similar temporal changes in saccade speed, blink rate, and pupil size as fatigue developed, suggesting that older adults adapt their scanning style but preserve the capacity to regulate fatigue-related changes in gaze dynamics over time. In older drivers specifically, Biernacki and Lewkowicz (2021) tested 44 older male drivers (65–88 years) in a simulator with a central *n*-back plus peripheral detection task and observed that higher load prolonged fixations and saccades—evidence for slower attentional shifting—with effects moderated by personality (greater slowing in those higher on neuroticism).

The possibility of physiological compensation has theoretical support from cognitive aging research. The Compensation-Related Utilization of Neural Circuits Hypothesis (CRUNCH) (Reuter-Lorenz and Cappell, 2008) posits that older adults recruit additional neural resources to maintain performance levels. Originally demonstrated through brain imaging showing older adults engaging additional cortical regions at high task loads, parallel findings in physiological measures have recently emerged. Kim et al. (2023) showed that healthy older adults exhibit larger task-evoked pupil dilations than younger adults under high cognitive load conditions, indicating greater effort mobilization with age. This compensatory pattern extends to fatigue-related oculomotor control: across three studies spanning driving-related tasks with younger and older adults (Experiment 1: 15 younger, 14 older), simulator tasks, and online visual tasks, Hu and Lodewijks (2021) reported systematic shifts in saccade metrics, fixation distributions, and other oculomotor features as fatigue accumulated. While several features generalized across age groups, others showed age-contingent patterns, indicating that temporal 'motifs' of fatigue in oculomotor control are not identical in younger and older populations, further supporting the notion of age-specific physiological adaptation strategies.

2.2.3. Behavioral adaptations and interventions

Age-related scanning can be modified through training, with gains persisting beyond immediate practice. In a controlled on-road intervention, Romoser (2013) trained older drivers to execute critical side glances when approaching and turning at intersections. Post-training, participants made significantly more secondary glances to potential threat zones, and these improvements were retained at a multi-week follow-up, indicating durable changes in scanning strategy rather than short-lived compliance. Complementing behavioral work, population-level sensory assessments identify plausible constraints on exploratory scanning: in the Swedish cohort of 109 older drivers (mean age ≈ 71 years; 59 men, 50 women), Thorslund et al. (2019) documented the prevalence and combinations of vision and hearing deficits relevant to driving; such sensory limitations provide mechanistic context for age-related reductions in peripheral sampling and increased gaze centralization reported elsewhere.

Evidence of adaptation extends to everyday driving behaviors. Derafshi et al. (2024) investigated whether cognitive impairment alters driving patterns in older adults using in-vehicle GPS loggers. Following 246 drivers over two years (230 cognitively normal, 16 with mild cognitive impairment), they found that impaired drivers relied on fewer, more familiar routes, consistent with reduced navigational flexibility. Nevertheless, both groups maintained driving independence, suggesting that route simplification helps preserve mobility in the face of early cognitive decline. Similarly, Thorslund et al. (2019) found that despite more frequent self-reported difficulties among hearing-impaired older drivers, hearing impairment was not associated with greater crash involvement. Their study of 109 older Swedish drivers revealed that drivers adopted compensatory behaviors such as increased mirror use and avoidance of challenging conditions. Notably, while spectacles were used 95% of the time when driving, hearing aids were only used 57% of the time, despite both measured and experienced benefits of hearing aids being significantly associated with usage patterns.

These adaptation strategies extend to automated driving contexts, where age-related differences become particularly salient. Zhou et al. (2022) used eye-tracking features in controlled takeover scenarios (33 participants; 32 scenarios) to predict situation awareness (SA) in real time, achieving actionable accuracy in the transition window. Age differences in takeover behavior are mixed and appear sensitive to warning time, scenario complexity, and secondary-task engagement: in a simulator with conditional automation, Clark and Feng (2017) compared 18 older (62–81 years) and 17 younger (18–35 years) drivers and found similar takeover reaction times given adequate warnings (4.5–7.5 s), though older drivers engaged far less in non-driving tasks during automation and tended toward more abrupt, conservative control inputs upon resuming control. Review work emphasizes gaze and pupil as high-value channels in deployable distraction/monitoring systems (especially when fused with other physiology and modeled temporally) (Michelaraki et al., 2023), while Merat and de Waard (2014) underscore that attention, SA, and transition management remain central human-factors concerns as automation advances.

2.3. Computational approaches and temporal analysis

2.3.1. Classification methods in driver monitoring

Recent reviews provide systematic evidence about the current state of classification approaches in driver monitoring research. Koay et al. (2022) reviewed studies from 2014–2021, documenting that binary classification dominates the field, typically distinguishing distracted versus normal, drowsy versus alert, or high versus low workload states. Their analysis revealed that most studies employ supervised classification with short sliding windows of ocular features, using traditional machine learning approaches (SVM, Random Forest, XGBoost) or increasingly, deep learning architectures (CNN/LSTM). These systems routinely achieved 85–95% accuracy in controlled simulator settings, though accuracy often declined in person-independent and naturalistic evaluations.

Early work on predicting driver intentions demonstrated the potential of physiological pattern analysis. Doshi and Trivedi (2009) developed dynamic Bayesian models combining gaze direction, head pose, and vehicle dynamics using data from 10 participants (ages 22–55) in an instrumented vehicle. Their model achieved 92.4% accuracy in predicting lane changes 3 s before execution, demonstrating that visual behavior patterns can anticipate driver decisions and laying groundwork for predictive monitoring systems. Building on the potential of

ocular features alone, Rahman et al. (2021) developed models using 13 gaze features including fixation duration, saccade amplitude, and blink rate from 33 participants (mean age 29.4 years). Their support vector machine classifier achieved 92% accuracy in distinguishing high from low cognitive load, surpassing typical binary classification accuracies of 80–90% reported in their literature review, and suggesting that eye tracking alone might suffice for physiological assessment.

Michelaraki et al. (2023) provide a PRISMA-style review of real-time distraction monitoring with an emphasis on ocular signals. Across the retained studies, most systems frame detection as a binary task (e.g., attentive vs. distracted, low vs. high workload) using short, sliding-window features from eye trackers—fixation and saccade rates, blink metrics, eyes-off-road time—and pupillometry (mean pupil diameter and dilation dynamics). When moving beyond binary to multiclass recognition (e.g., assigning gaze to zones such as road, mirrors, or dashboard, or distinguishing multiple distraction types), reported accuracy generally drops. Methodologically, classical machine-learning pipelines (e.g., support vector machines, random forests) remain common, with deep models mainly used for eye-tracking processing; person-independent validation and longer on-road evaluations are still relatively rare. The review also notes that explicit age stratification is rare—leaving open how gaze and pupil dynamics, and hence model performance, vary across younger and older drivers in real-time settings.

More complex event detection was attempted in naturalistic driving data from the Second Strategic Highway Research Program Naturalistic Driving Study (SHRP 2 NDS) dataset with 1820 crashes, 6848 near-crashes, and 59,997 normal driving segments (Shi et al., 2022). Models combining temporal gaze patterns with contextual features (e.g., vehicle speed, headway, roadway level of service) used a hybrid architecture in which a 1D CNN extracted short-horizon temporal descriptors that were subsequently classified with XGBoost; the task was multiclass (crash, near-crash, baseline), with class imbalance handled via sampling/weighting. Overall accuracy of up to 97.5% was reported for crash/near-crash/baseline discrimination at the segment level, though confusions between crash and near-crash were more common than confusions with baseline (the precision and recall for crashes were 84.7%, and 71.3% respectively), underscoring the difficulty of fine-grained risk stratification beyond binary detection (Shi et al., 2022).

Beyond simulator studies, field validation demonstrates the feasibility of real-world physiological monitoring. In two real-world highway studies, Solovey et al. (2014) classified driver workload online from multimodal physiology and vehicle performance (Study 1: N=20; Study 2: N=99). They framed a binary classification task (elevated vs. baseline workload) using short sliding windows and supervised learners, with signals including ECG-derived heart rate/variability, respiration (rate/depth), and electrodermal activity alongside vehicle telemetry (e.g., speed, lane position/variability, steering/brake activity). Across within-subject and leave-one-driver-out protocols, classifiers such as SVM/k-NN/Naive Bayes reached accuracies up to ∼90% in the field, with consistent gains when physiology was fused with performance features rather than using vehicle data alone; feature-importance analyses highlighted pupil/EDA and steering/lane-keeping variability among top contributors (Solovey et al., 2014). These results confirm that combining physiological measures with performance yields higher accuracy than performance alone and that cross-driver generalization is feasible even in uncontrolled environments.

From a methodological perspective, Sharma and Chakraborty (2024) conducted a systematic review of 154 driving studies (up to mid-2023) documenting a split between feature-based classifiers using fixation and saccade metrics versus deep architectures learning spatiotemporal representations directly from eye-tracking data. While deep models capture richer temporal dynamics, feature-based approaches remain attractive for transparency and lighter computation. They detail the most commonly used ocular measures for driver monitoring: fixation/dwell time, time-to-first-fixation, and fixation count; saccade metrics including count, amplitude, duration, and scanpaths; glance duration/frequency and transitions across gaze zones; pupil size and dilation dynamics; blink rate; and entropy-style summaries quantifying dispersion and sequencing of gaze. Their review highlights practical deployment challenges—models trained in simulators often show substantial degradation when applied to real driving (one I3D-based model's accuracy dropped from 85.7% to 46.6%), person-independent validation remains rare, and explicit age stratification is uncommon. Regarding age effects, they note that older drivers tend to scan less laterally at intersections, rely more on lane markings, and make fewer mirror/blind-spot glances during lane changes—patterns directly relevant to gaze- and pupil-based monitoring of safety-critical maneuvers, though these effects appear context-dependent and require studies explicitly designed to test age differences.

2.3.2. Time-series analysis and pattern discovery

While the aforementioned studies have established clear behavioral and physiological markers of driving states and age-related differences, they predominantly rely on aggregated features over fixed or short sliding time windows. This approach may miss temporal dynamics that could provide additional discriminative power or reveal compensatory patterns.

Statistical feature extraction from fixed temporal windows has dominated physiological analysis in driving research. Hogervorst et al. (2014) compared multiple modalities using data from 14 participants performing simulated air traffic control tasks, finding that eye-derived features achieved 86% classification accuracy, comparable to EEG (87%) while using simpler sensors. Window length selection critically impacts classification performance, as demonstrated by Meteier et al. (2021) who systematically evaluated windows from 30 s to 5 minutes using 90 drivers' data. Performance peaked at 2–4 minute windows (95% accuracy), revealing an optimal temporal scale for capturing relevant physiological dynamics.

Wang et al. (2024) conducted a systematic meta-analysis addressing inconsistent findings about physiological markers of cognitive load in driving. Using a random-effects model, they categorized metrics into four sensitivity tiers: high-resolution indicators (pupil size, heart rate, skin conductance) that reliably differentiated low, medium, and high load; sensitive-to-low metrics (e.g., EEG theta at Fp1) distinguishing only low from medium load; low-resolution measures (e.g., ECG power, blink rate, respiration rate) effective only for extreme comparisons; and inconsistent measures (e.g., EEG theta at Fp2). Moderator analyses revealed that task characteristics (repeating vs. matching vs. counting n-back variants), stimulus modality, driving automation level, and demographic factors systematically modulated metric sensitivity. These findings provide evidence-based guidance for selecting metrics when designing monitoring systems, balancing sensitivity with task demands and participant variability.

To advance beyond these traditional approaches, several research directions have emerged. Tao et al. (2024) introduced a comprehensive multimodal physiological driving dataset (MPDB) to facilitate advanced time-series analyses in driving behavior research. The MPDB includes high-density EEG (59 channels), ECG, EMG, GSR, and eye-tracking data collected from 35 drivers in a simulator, labeled across five representative driving states (smooth driving, acceleration, deceleration, lane changing, turning). In their baseline analysis, Tao et al. demonstrated the dataset's utility by training both traditional classifiers (LDA) and deep neural networks (MMPNet, EEGNet) to predict driving maneuvers from synchronized physiological signals. Classification performance reached moderate accuracy (on the order of 60–65%), which they attributed to the complexity and variability of physiological responses, emphasizing that such signals may not directly encode discrete maneuvers in a one-to-one mapping.

Complementing supervised learning approaches, unsupervised pattern discovery methods provide another direction for extracting meaningful temporal structure from physiological signals. The Matrix Profile algorithm enables discovery of recurring patterns without prior pattern

definitions, with Law's STUMPY implementation (Law, 2019) providing efficient computation for large time-series. Silva and Henriques (2021) demonstrated this potential with their TripMD system, which analyzed naturalistic vehicle telematics data sampled at high frequency. Using a matrix profile-based framework, they automatically extracted recurrent driving motifs such as acceleration, braking, and turning, without requiring labeled training data. The authors showed that TripMD was able to distinguish individual driving styles and identify consistent maneuver patterns that aligned with driver behavior. Importantly, the system proved more adaptable to small fluctuations than supervised models, highlighting the robustness of motif-based unsupervised discovery in capturing meaningful sequential structure directly from raw vehicle sensor streams. Outside the driving domain, Kraljevska et al. (2025) showed how motif discovery can capture predictive temporal patterns in psychiatric EEG using matrix profile-based k-Motiflets, illustrating methodological potential relevant to driver monitoring contexts.

Temporal structure in control signals can anticipate performance changes before overt failures. In a 60-minute simulator drive with $N=17$, Zhang et al. (2022) quantified steering entropy – operationalized as aggregated prediction error of near-future steering angle from a short-horizon Taylor expansion – and found that entropy increased after ~15 minutes while lane-keeping variability (SDLP) did not degrade until ~30 minutes. This temporal dissociation suggests that continuously computed indices (like steering entropy) can provide an early warning of state change. Such control-signal markers conceptually complement time-resolved ocular measures (e.g., pupil dynamics, gaze dispersion) in monitoring pipelines.

For handling irregular sampling common in real-world physiological monitoring, Ren et al. (2024) developed a deep temporal interpolation and clustering (dTIC) network. Using data from over 43,000 patients, they jointly interpolate irregular measurements, learn latent temporal embeddings, and apply unsupervised clustering. Results revealed four reproducible patient phenotypes with distinct comorbidity profiles, demonstrating how unsupervised time-series methods can recover structured physiological dynamics often obscured by conventional preprocessing. Deep learning approaches have also shown promise for automatic feature learning, with Vortmann et al. (2021) converting gaze and pupil sequences into 2D spectrograms for CNN classification, achieving 89% accuracy on workload tasks compared to 81% with traditional features. In their study with 28 participants, sequential ocular signals were transformed into time-frequency spectrograms and used as image inputs to a convolutional neural network; performance was benchmarked against a hand-crafted feature baseline built from conventional gaze and pupil metrics, highlighting the benefit of time-frequency representations for capturing discriminative temporal structure. However, the interpretability challenge persists – understanding which temporal patterns drive predictions in neural networks requires additional analysis, unlike tree-based models that readily support explainability methods such as SHAP.

These works show that age stratifies temporal response organization in driving and that time-resolved physiological/behavioral signals – especially gaze and pupil – are strong indicators of workload, situation awareness, fatigue/recovery, and impending performance degradation. Broader syntheses on aging and performance situate these observations within known changes in perception, attention, and executive control that bear directly on visual sampling and regulation in dynamic tasks (Charness, 2008; Anstey et al., 2005).

2.4. Sensor capabilities and constraints

Practical deployment of physiological monitoring faces sensor-related constraints. The introduction of affordable depth cameras expanded monitoring capabilities, though with limitations – Sun et al. (2015) showed that while Kinect sensors could estimate gaze direction through head pose, accuracy remained inferior to dedicated eye trackers. González-Ortega et al. (2017) confirmed this finding, achieving only 67% gaze zone accuracy from Kinect head pose estimation versus 98% from direct eye tracking. This substantial accuracy gap highlights the trade-off between sensor cost and measurement precision.

Multimodal sensor fusion has been proposed to improve robustness. Walker et al. (2019) combined gaze metrics with electrodermal activity to assess automation trust, finding that low trust manifested through both increased visual scanning and heightened physiological arousal. Similarly, Rauh et al. (2020) fused ocular metrics with heart rate variability for real-world fatigue detection. He et al. (2022) extended this line of work with a simulator study of 33 drivers, comparing five classifiers across eye-tracking, heart rate, and galvanic skin response modalities. While eye-tracking alone yielded the highest single-modality accuracy (up to 97.8% with Random Forests), adding physiological channels consistently boosted performance—particularly galvanic skin response, which improved accuracy by up to 29% relative to eye-only models.

2.5. Research gaps

While extensive literature exists on physiological monitoring during driving, several critical gaps remain. Meta-reviews consistently document that binary classification approaches dominate the field (Koay et al., 2022; Michelaraki et al., 2023), with studies typically examining states such as high versus low workload or distracted versus alert conditions. Multi-class discrimination between specific driving events has received limited attention beyond gaze-zone recognition or distraction type categorization.

Additionally, age-related physiological adaptations in driving contexts remain understudied. While behavioral decline in older drivers is well-documented (Maltz and Shinar, 1999; Underwood, 2007), potential compensatory mechanisms whereby older drivers might enhance certain physiological responses to maintain performance have not been systematically investigated, despite theoretical support from cognitive aging research.

Methodological approaches in existing studies typically employ single analytical frameworks, missing opportunities for convergent validation across complementary methods. Furthermore, the relationship between physiological signal characteristics and their behavioral relevance remains unexamined, particularly whether stronger responses necessarily provide better event discrimination or if subtle temporal patterns might be more informative.

This study examines gaze and pupil dynamics during simulated driving to explore multiclass event discrimination and potential age-related patterns using complementary analytical approaches.

3. Materials and methods

3.1. Participants

A total of 27 participants took part in the study: 12 younger adults (ages 20–29; $M = 23.1$, $SD = 2.7$; 7 female / 5 male) and 15 older adults (ages 61–74; $M = 66.7$, $SD = 4.0$; 7 female / 8 male). These age-group definitions are consistent with established age stratifications in driving and visual attention research (Maltz and Shinar, 1999; Cooper et al., 2020). Among the older adults, seven completed the study using a minimal head-up display (HUD) variant of the simulator that provided basic navigation information (4 male / 3 female). The HUD was enabled for this subset as an initial exploration of whether navigation assistance might benefit older drivers; however, HUD effects were not the primary focus of this study. Because HUD presence could represent a potential confound, its influence on physiological responses was evaluated through sensitivity analyses reported in Section 4.1.1.

For sensitivity analysis, an ablation subset was defined by excluding HUD users. This no-HUD set comprised 20 participants: 12 younger adults (7 female / 5 male) and 8 older adults (4 female / 4 male). Throughout the manuscript, results are reported for both the combined dataset ($n = 27$) to maximize statistical power and the ablation

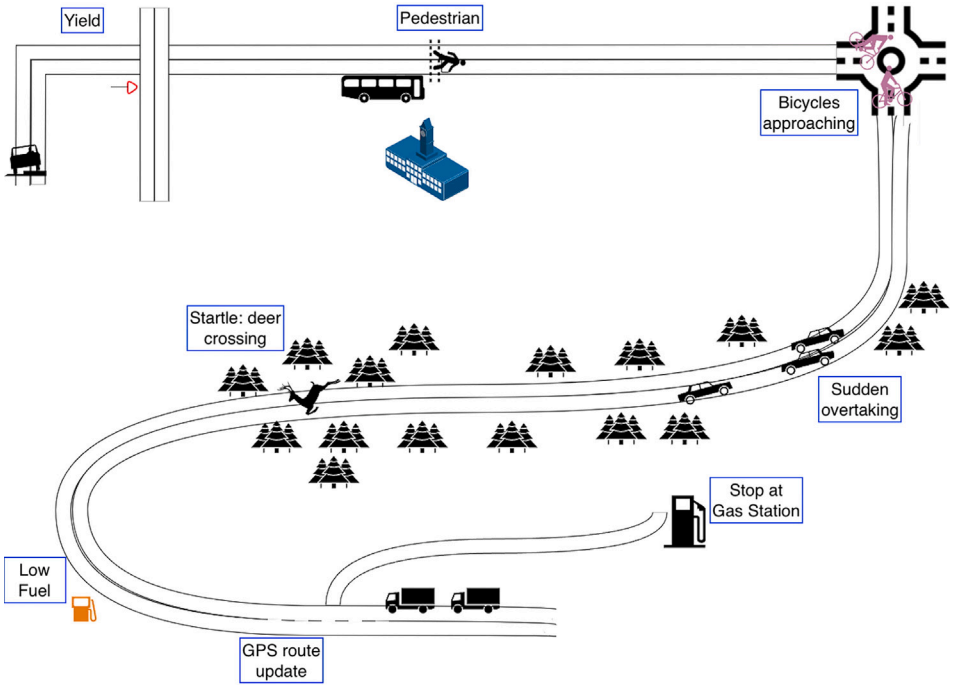


Fig. 1. Schematic overview of the standardized driving scenario. Participants encountered a fixed sequence of critical events interleaved with baseline driving periods: pedestrian crossing, bicycle approach and sudden stop, vehicle overtaking, deer crossing (startle), fuel warning, GPS update, and gas station stop. Event ordering was fixed, while inter-event durations varied with driving speed.

subset ($n = 20$) to ensure that display modality differences do not confound the main findings. The consistency of results across both datasets strengthens the generalizability of the conclusions.

The sample size of 27 participants aligns with established driving simulator studies in the field, where samples of 20–30 participants are standard for physiological monitoring research (Mehler et al., 2012; Čegovnik et al., 2018; Nilsson et al., 2018). Sensitivity analysis indicated that with $N = 27$ participants (12 younger, 15 older) across the four analyzed events, the design was powered to detect large effects (minimum detectable Cohen's $d \approx 1.13$ for group comparisons; $f \approx 0.69$, $\eta^2 \approx 0.48$ for event comparisons). In the no-HUD subset ($N = 20$; 12 younger, 8 older), the detectable group effect was larger ($d \approx 1.35$). These calculations confirm that while our sample was adequate for detecting the large effect sizes observed ($\eta^2 > 0.5$ for primary measures), subtle age-related patterns would require larger samples.

All participants held valid driver's licenses, had normal or corrected-to-normal vision, and reported no neurological or musculoskeletal impairments.

3.2. Experimental setup

The experiment utilized a compact motion-based driving simulator at Faculty of Electrical Engineering, University of Ljubljana, featuring authentic vehicle components (seat, steering wheel, pedals, dashboard) and three 48-inch curved Samsung displays providing a 135° horizontal field of view. The simulator is based on a 4-degree-of-freedom motion platform (pitch, roll, heave, and surge), providing synchronized vestibular feedback during acceleration, braking, and turning. The 25.5-kilometer simulated route lasted 17–22 minutes depending on driving speed, encompassing suburban roads, rural segments with wildlife, and complex urban traffic conditions.

Two synchronized physiological recording systems captured participant responses. Depth signals were recorded using an Intel RealSense D435 depth camera (30 Hz) (IntelRealSense, 2020) mounted behind the dashboard. Post-hoc inspection revealed that the minimum distance field consistently tracked the hands-steering region (0.15–0.17 m range) rather than the intended torso, limiting body movement analysis to steering proximity measures. Eye tracking was performed using Tobii Pro Glasses 2 (50 Hz) (Tobii, 2020), recording normalized binocular gaze position across the unified display area, with coordinates normalized to [0,1] where (0.5, 0.5) represents the center of the combined display field, and continuous binocular pupil diameter.

Participants requiring vision correction used Tobii Pro's manufacturer-provided prescription lens inserts, which are magnetically integrated into the Tobii Pro Glasses 2 head unit and replace the use of external eyeglasses. The eye tracker was calibrated individually for each participant while wearing the prescription lenses prior to the test drive, ensuring that gaze mapping accounted for the participant's optical configuration.

3.3. Experimental protocol

Participants completed a standardized driving scenario with a fixed event sequence (see Fig. 1 and Table 1). All participants completed a brief familiarization drive to acclimate to the simulator environment. No participants reported motion sickness or discomfort during the test or experimental sessions. After a 2-minute adaptation period, a 30-second baseline (`baseline_main`) was extracted, with recordings truncated 15 s after the final event. While event ordering remained fixed, inter-event durations varied by driving speed (baseline SDs: 15–30 s). The participants were exposed to several driving situations, interspersed with baseline periods, that required different levels of engagement and situational awareness to be successfully resolved: pedestrian crossing, bicycle approach, sudden bicycle stop, vehicle overtaking, deer crossing (startle event), fuel warning, intersection waiting, GPS update, and gas station approach. Representative video recording (from Tobii glasses) of the four analyzed events, captured from the driver's perspective, is provided in the data repository (see Data Availability Section).

3.4. Data processing

Signals were preprocessed using custom Python scripts with depth values outside the 1st–99th percentiles interpolated, gaze coordinates

Table 1
Standardized driving sequence with mean duration per segment (n = 20 participants).

Segment	Description	Duration (s)
baseline_main	Initial uneventful driving	96.68
Pedestrian	Pedestrian crossing	4.68
baseline_1	Inter-event baseline	195.69
Bicycle	Bicycle approaches	14.15
StopBicycle	Bicyclist stops abruptly	6.00
baseline_2	Inter-event baseline	244.80
OvertakingVehicleAlert	Sudden overtaking	3.34
baseline_3	Inter-event baseline	140.39
DeerAlert	Startle: deer crossing	18.28
baseline_4	Inter-event baseline	280.36
LowGas	Dashboard fuel warning	16.31
baseline_5	Inter-event baseline	38.64
GPS	GPS route update	8.62
StopAtGasStation	Refueling instruction	23.00

converted to angular deviations from center, and pupil channels averaged and smoothed. Pupil diameter values outside the physiologically plausible range (2–8 mm) were excluded prior to analysis, ensuring that spurious samples due to tracking noise or optical artifacts did not contribute to downstream features. Approximately 4.8% of pupil data remained missing after cleaning due to tracking loss or blinks. Raw pupil diameter values were retained for absolute group-level comparisons, and z-scores normalized within-participant relative to each event were used for event-level analyses and modeling.

Six primary features were extracted: mean pupil diameter (mm), mean steering proximity from depth sensor (m), and four gaze metrics including horizontal deviation, vertical deviation, total angular distance, and normalized position from center (Table 2). Additional range metrics were computed for each feature to capture variability within events. Despite the depth sensor's misalignment, the data were retained for exploratory analysis as statistical tests revealed significant between-group and between-event differences, suggesting potential behavioral information even from this unintended measurement region.

3.5. Statistical analysis

Non-parametric tests were employed due to violations of normality (Shapiro–Wilk) and variance homogeneity (Levene's test). Mann–Whitney U tests with rank-biserial correlation assessed between-group comparisons, while Kruskal–Wallis H tests with η^2 effect sizes evaluated between-event differences. Post-hoc pairwise comparisons used Mann–Whitney U tests with Holm correction for multiple comparisons, and statistical power was computed for significant contrasts.

To distinguish genuine behavioral variation from measurement error, a base-rate error analysis calculated the proportion of observations exceeding sensor noise thresholds: ±3 mm for depth (2% accuracy at 0.15–0.17 m operating range) (IntelRealSense, 2020), ±0.3° for gaze, and physiologically implausible pupil diameter values outside the 2–8 mm range (Hooge et al., 2023). The gaze threshold of ±0.3° was selected as a measurement precision floor based on published evaluations of wearable eye-tracker performance under head and body motion (Niehorster et al., 2024), ensuring that observations exceeding this threshold reflect genuine behavioral variation rather than sensor noise. This approach complements significance testing by ensuring that observed effects exceed expected sensor noise.

To assess global structure in motif-derived spaces, Permutational Multivariate Analysis of Variance (PERMANOVA) was applied to distance matrices using the `adonis2` function from the R `vegan` package. Two permutation designs were evaluated: (i) unstratified, with age labels permuted freely across all windows, and (ii) stratified by event×signal, constraining label permutations within each panel to control for panel-specific composition. For each design, 1000 permutations were performed, yielding proportions of variance explained (R^2) and permutation p-values. Benjamini–Hochberg false discovery rate (BH-FDR) correction was applied across global tests. Panel-wise PERMANOVA was additionally conducted within each event×signal combination meeting minimum sample requirements (at least seven windows per age group), with BH-FDR correction applied across all panel-level tests.

Table 2
Physiological features with definitions.

Feature	Unit	Definition
pupil_mean	mm	Mean cleaned pupil diameter
depth_mean	m	Mean steering proximity (minimum depth distance)
gaze_hdev_mean	degrees	Mean horizontal deviation from center
gaze_vdev_mean	degrees	Mean vertical deviation from center
gaze_distance_mean	degrees	Total angular deviation (Euclidean)
gaze_position_mean	0–1	Normalized eccentricity from center

3.6. Machine learning analysis

For machine learning classification, signals were segmented into fixed 9-second windows (450 samples at 50 Hz) aligned to event onset, a duration necessary for computing temporal and frequency-domain features while preserving event specificity. The 9-second window duration was selected to capture temporal and frequency-domain features of physiological responses, including pupil dilation dynamics (typically 2–5 s) and gaze scanning patterns (3–9 s for complex visual search). This duration balances temporal resolution with the need for sufficient data to compute spectral features (0.1–2.0 Hz bands), while maintaining event specificity. Additionally, 9 s represents the longest fixed window applicable uniformly across all participant–event instances: event durations varied across participants due to differences in driving speed and maneuver timing, with the shortest observed segment being 9.26 s (participant 261, LowGas). Using event-specific durations would confound time-dependent features (e.g., AUC, variability) with duration effects. Sensitivity analysis confirmed that classification performance increased monotonically with window length (3 s: macro-F1 = 0.639; 6 s: macro-F1 = 0.732; 9 s: macro-F1 = 0.824); full-event analysis (macro-F1 = 0.842) is reported as a robustness check in Supplementary Table S6. This constraint restricted analysis to four events exceeding 9 s: `Bicycle` (14.15 s), `DeerAlert` (18.28 s), `LowGas` (16.31 s), and `StopAtGasStation` (23.00 s); see Table 1.

From each window, 68 features were computed across six physiological signals capturing statistical properties (mean, median, standard deviation, minimum, maximum, range), temporal dynamics (area under curve, linear slope, 1-second delta changes), and frequency characteristics (FFT peak frequency and spectral band power for 0.1–0.5 Hz and 0.5–2.0 Hz bands). Signal-specific features included peak amplitude and latency for pupil responses, and variance, dispersion, and rapid change counts for gaze position.

Two classification tasks were evaluated using gradient boosted decision trees (XGBoost) with leave-one-participant-out cross-validation (LOPO-CV): multiclass discrimination between the four events, and binary classification of each event versus baseline driving. Feature selection for multiclass classification employed aggregate gain ranking across all LOPO folds, with systematic evaluation of top-K features (K = 5, 7, 10, 12, 15) selecting K = 7 based on macro-F1 score optimization and reduced feature set. Binary classification used fold-specific feature selection retaining features with mutual information exceeding 0.05. Model hyperparameters remained fixed across experiments with 400 estimators, learning rate of 0.05, maximum depth of 5, subsample ratio of 0.9, and balanced class weights addressing event frequency imbalances. Performance evaluation included balanced accuracy, F1 scores, and ROC-AUC metrics.

SHAP (SHapley Additive exPlanations) values (Lundberg and Lee, 2017) quantified feature contributions to event classification, computed

using TreeExplainer for each held-out participant with their fold-specific model to ensure no data leakage. Feature importance was assessed through mean absolute SHAP values both globally and stratified by event type and age group. SHAP analysis provided post-hoc interpretability for the features selected through gain ranking, revealing how each feature contributed to event discrimination rather than validating the selection itself. This distinction is important: feature selection was based on XGBoost's gain metric across LOPO folds, while SHAP values subsequently quantified how these selected features influenced classification decisions.

3.7. Conceptual framework for temporal pattern analysis

The analysis distinguished three hierarchical levels of temporal pattern abstraction to capture both individual behavioral strategies and collective response prototypes:

Motifs: individual behavioral strategies. Motifs represent recurring temporal patterns within individual participants' physiological signals, revealing intra-individual structure—systematic response strategies reproduced under similar conditions. Unlike aggregated features that discard temporal information, motifs preserve dynamic response shapes, capturing how individuals structure their physiological responses over time (e.g., repeated gaze shifts between road and dashboard, recurrent pupil dilations during monitoring).

Archetypes: collective response prototypes. Archetypes are canonical temporal patterns shared across participants, derived through k-means clustering of pooled motif windows. These centroidal shapes capture inter-individual regularities—common behavioral prototypes generalizing beyond single participants (e.g., prototypical angular gaze excursion during hazard response, characteristic pupil drift during dashboard monitoring). Archetypes provide a library of shared strategies for interpreting collective temporal dynamics.

Prototypes: statistical representatives. For statistical inference, prototypes serve as group representatives within each event×signal panel. Each age group is represented by its medoid—the observed window minimizing average DTW distance to other group members. Unlike synthetic archetypes, prototypes remain actual observed patterns, enabling direct statistical comparison while maintaining interpretability.

This framework preserves temporal structure lost in traditional aggregate measures, revealing how physiological responses manifest as structured behavioral strategies (motifs) that converge into collective patterns (archetypes) and support statistical inference through representative exemplars (prototypes).

3.8. Time-series motif discovery

Motif detection. Time-series motifs were identified using the matrix profile algorithm (Law, 2019), which computes z-normalized Euclidean distances between all subsequence pairs. Subsequences of m = 100 samples (2.0 s at 50 Hz) balanced capturing physiological transients against oversmoothing. For each participant-event combination across five signals (gaze horizontal, vertical, angular, position, and pupil), the pipeline: (i) computed the matrix profile, (ii) selected up to three motif seeds with 50-sample exclusion zones, (iii) expanded seeds to all matches within matrix-profile radius, and (iv) merged overlapping detections into non-overlapping bursts. Three metrics characterized motifs: frequency (bursts/s), coverage (fraction of samples within spans), and amplitude (z-normalized range).

Archetype construction. Group-level temporal structure was evaluated through k-means clustering (k = 5) of resampled motif spans, creating centroidal archetypes summarizing recurring patterns across participants. Intersection-over-union (IoU) validated overlap between archetype and detected motif spans. Additionally, motif distance (MPdist) linked temporal patterns to classification relevance by quantifying similarity between participants' motifs and those with highest SHAP importance.

Prototype-based group comparisons. Age-related differences were quantified using medoid-based prototypes within each event×signal panel. Between-group separation was defined as $\Delta_{\mathrm{DTW}} = \mathrm{DTW}(\mathrm{medoid}_{\mathrm{young}}, \mathrm{medoid}_{\mathrm{old}})$. Within-group dispersion ($W_{\mathrm{young}}, W_{\mathrm{old}}$) measured mean DTW distance from individuals to their group medoid.

Permutation testing (2,000 iterations per panel) was used to assess statistical significance, with BH–FDR correction applied across panels. Analyses required at least ≥ 7 windows per age group. Dispersion contrasts ($W_{\mathrm{old}} - W_{\mathrm{young}}$) were used to test within-group homogeneity differences. Optional DTW barycenters (DBAs) were included for visualization only; all statistical inference relied on medoid-based metrics.

Analyses were conducted in Python 3.10 using NumPy (v1.24.3) (Harris et al., 2020), pandas (v2.0.3) (McKinney et al., 2010), SciPy (v1.10.1) (Virtanen et al., 2019), Pingouin (v0.5.3) (Vallat, 2018), Matplotlib (v3.7.2) (Hunter, 2007), seaborn (v0.12.2) (Waskom, 2021), SHAP (v0.42.1) (Lundberg and Lee, 2017), STUMPY (v1.12.0) (Law, 2019), and XGBoost (v1.7.6) (Chen and Guestrin, 2016).

4. Results

Statistical analyses were conducted across all events to provide comprehensive characterization of physiological responses to diverse driving scenarios, whereas machine learning and motif analyses were restricted to four events exceeding 9 s in duration (Bicycle, DeerAlert, LowGas, StopAtGasStation), as specified in Methods Section 3.6. This dual approach balances broad behavioral characterization with methodological requirements for temporal feature extraction.

4.1. Statistical analysis

4.1.1. Combined and ablation set comparison

Before presenting the main findings, the study first examined whether inclusion of HUD users among older participants introduced systematic biases. Statistical analyses were therefore conducted on both the ablation set (20 non-HUD participants) and the combined set (all 27 participants).

Mann–Whitney tests comparing HUD and non-HUD subgroups within the older adult sample revealed no significant differences in any physiological measure. Pupil diameter showed negligible effects ($p_{\mathrm{Holm}} = 0.46$, $|r| < 0.15$), as did normalized pupil diameter ($p_{\mathrm{Holm}} = 0.62$), horizontal gaze deviation ($p_{\mathrm{Holm}} = 0.55$), vertical gaze deviation ($p_{\mathrm{Holm}} = 0.41$), and total gaze distance ($p_{\mathrm{Holm}} = 0.48$). These small effect sizes persisted when examining individual events: no HUD-related differences survived correction across any of the four modeling events ($p_{\mathrm{Holm}} > 0.20$ for all comparisons). For instance, horizontal gaze deviation during `StopAtGasStation` showed virtually identical medians between HUD users (5.63°) and non-HUD users (5.72°), exemplifying the minimal impact of display modality on physiological responses.

This absence of systematic HUD effects justifies pooling all 27 participants for primary analyses, thereby increasing statistical power while maintaining the original non-HUD cohort for sensitivity checks. Raincloud plots illustrating these overlapping distributions are provided in Supplementary Figure S1, confirming that HUD usage does not meaningfully alter gaze or pupil behavior patterns in our older adult sample.

4.1.2. Signal reliability assessment

Base-rate error analysis established the reliability of our physiological measurements by comparing observed values against sensor noise thresholds defined in Section 3.5. Pupil diameter and gaze measures consistently exceeded their respective noise floors, with 94% of pupil observations and 87–92% of gaze observations surpassing measurement precision limits. This high signal-to-noise ratio confirmed that both

modalities captured genuine behavioral variation rather than sensor artifacts. In contrast, depth measurements from the misaligned camera showed poor reliability, leading us to exclude this signal from further analysis despite its initial statistical significance.

Physiological stability across the driving session was verified through baseline comparisons. Most measures remained stable throughout the experiment, with only modest deviations observed late in the session. Specifically, horizontal gaze showed a small but significant rightward shift during `baseline_5` compared to the initial baseline ($p < 0.001$; $\text{Mdn}_{\text{main}} = -0.162°$ vs. $\text{Mdn}_{\text{int}} = 1.434°$), while total gaze distance exhibited a non-significant trend toward increased values ($p = 0.065$; 6.687 vs. 7.831). These late-session shifts likely reflect anticipatory adjustments as participants approached the final driving segments, but their modest magnitude confirms that baseline drift did not confound event-specific responses.

4.1.3. Event-wise differences

Driving events elicited robust modulation of physiological responses, as revealed by Kruskal–Wallis H tests across both datasets (Table 3). The pattern of results was consistent between the combined and ablation sets, with effect sizes remaining large despite the addition of seven HUD users.

In the combined dataset ($n = 27$), normalized pupil variability emerged as the most sensitive metric ($\eta^2 = 0.973$, 95% CI [0.965, 0.982]), indicating that event-specific changes in pupil dynamics provide a robust discriminative signal. Gaze dispersion metrics followed closely, with vertical deviation range ($\eta^2 = 0.577$, 95% CI [0.523, 0.654]), position range ($\eta^2 = 0.560$, 95% CI [0.504, 0.642]), and total distance range ($\eta^2 = 0.554$, 95% CI [0.493, 0.643]) all showing substantial event-driven modulation. These large effect sizes indicate that visual scanning patterns adapt markedly to different driving demands, with events constraining or expanding the spatial extent of gaze exploration.

The ablation set ($n = 20$) showed nearly identical patterns, with normalized pupil variability maintaining an exceptionally large effect ($\eta^2 = 0.960$, 95% CI [0.949, 0.972]) and gaze metrics retaining their discriminative power. The consistency of these findings across both datasets demonstrates that event-specific physiological signatures are robust to minor variations in display technology and sample composition. All significant effects achieved high statistical power (>0.97), confirming that our sample size was adequate to detect these large behavioral modulations.

Post-hoc pairwise comparisons revealed which specific event pairs drove these omnibus effects. Statistical analyses across all recorded events identified significant contrasts between event types, though these patterns varied between datasets. The four events selected for machine learning analyses (Bicycle, DeerAlert, LowGas, StopAtGasStation) showed robust discrimination in both combined and ablation datasets, confirming their suitability for temporal modeling.

Table 3
Event-wise comparisons using Kruskal–Wallis H tests with Holm correction. Results for the combined dataset ($n = 27$) are shown above, and for the ablation dataset ($n = 20$, non-HUD only) below.

Combined (27 participants)				
Feature	p (Holm)	η^2	95% CI	Power
pupil_norm_std	< 0.001	0.973	[0.965, 0.982]	1.000
gaze_vdev_range	< 0.001	0.577	[0.523, 0.654]	1.000
gaze_position_range	< 0.001	0.560	[0.504, 0.642]	1.000
pupil_range	< 0.001	0.532	[0.493, 0.603]	1.000
gaze_distance_range	< 0.001	0.554	[0.493, 0.643]	1.000
Ablation (20 participants, non-HUD only)				
Feature	p (Holm)	η^2	95% CI	Power
pupil_norm_std	< 0.001	0.960	[0.949, 0.972]	1.000
gaze_vdev_range	< 0.001	0.571	[0.508, 0.648]	1.000
gaze_position_range	< 0.001	0.571	[0.510, 0.647]	1.000
pupil_range	< 0.001	0.498	[0.445, 0.578]	1.000
gaze_distance_range	< 0.001	0.582	[0.523, 0.660]	1.000

4.1.4. Age group differences

Age-related differences in physiological responses were examined using raincloud plots to visualize the full distributions alongside summary statistics (Fig. 2). These plots combine raw data points, kernel density estimates, and violin plots from the combined dataset (n = 27) to provide a comprehensive view of between-group differences while preserving information about within-group variability and potential outliers. For reference, dashed lines indicate the medians from the ablation set (n = 20, non-HUD only), allowing direct comparison between the expanded and original samples.

Visual inspection reveals clear separation between age groups across multiple measures, with the combined dataset distributions closely aligning with the ablation set medians. While raw pupil diameter shows pronounced age differences, these largely disappear after z-normalization within participants, indicating anatomical rather than functional disparities. In contrast, gaze dispersion metrics show clear age-related patterns, with younger participants exhibiting broader distributions for gaze measures, particularly for horizontal and angular deviation, suggesting more variable visual exploration strategies. The consistency of these gaze patterns across different driving events indicates fundamental age-related differences in visual scanning behavior rather than event-specific adaptations.

Statistical comparisons using Mann–Whitney U tests confirmed these visual observations (Table 4). In the combined dataset, raw pupil diameter showed a large age effect ($p < 0.001$, rank-biserial = −0.707, 95% CI [−0.783, −0.632]). However, when examining within-participant z-normalized pupil diameter, the effect magnitude dropped substantially (rank-biserial = −0.141, 95% CI [−0.307, 0.025]) and lost statistical significance ($p_{\text{Holm}} = 0.427$). This dramatic reduction confirms that raw pupil differences primarily reflect age-related changes in baseline eye physiology rather than differential task-related modulation. Both age groups showed comparable event-driven pupil responses when scaled to their individual baselines.

The non-significant difference in normalized pupil diameter between age groups ($p_{\text{Holm}} = 0.427$) indicates comparable event-driven pupil responses when scaled to individual baselines. However, subsequent motif analysis revealed systematic differences in temporal pupil dynamics (Section 4.3), suggesting that aggregate measures may mask temporal patterns that distinguish age groups.

Beyond pupillary measures, gaze behavior also differed systematically between age groups. Younger participants demonstrated greater gaze position variability ($p = 0.001$, rank-biserial = −0.238, 95% CI [−0.341, −0.120]), indicating more extensive spatial exploration of the visual field. Angular gaze deviation similarly showed age-related reduction ($p = 0.016$, rank-biserial = −0.197, 95% CI [−0.307, −0.092]), suggesting that older drivers maintain more constrained scanning patterns. These effects achieved near-perfect statistical power (0.974–0.997), confirming robust detection of moderate effect sizes.

The ablation dataset yielded consistent results with slightly attenuated effect magnitudes, as expected given the smaller sample size. Raw pupil diameter maintained a strong age effect (rank-biserial = −0.626), while normalized pupil diameter again showed no significant group difference. Gaze position variability remained significantly different between groups ($p = 0.002$, rank-biserial = −0.259), though the angular deviation difference did not remain statistically significant after Holm correction in this smaller sample. The convergence of findings across both datasets strengthens confidence that these age-related patterns reflect genuine behavioral differences rather than sampling artifacts.

4.1.5. Age-specific response patterns

To examine whether age groups showed differential sensitivity to specific driving events, we tested age-by-event interactions. The combined dataset revealed a single significant interaction: horizontal gaze

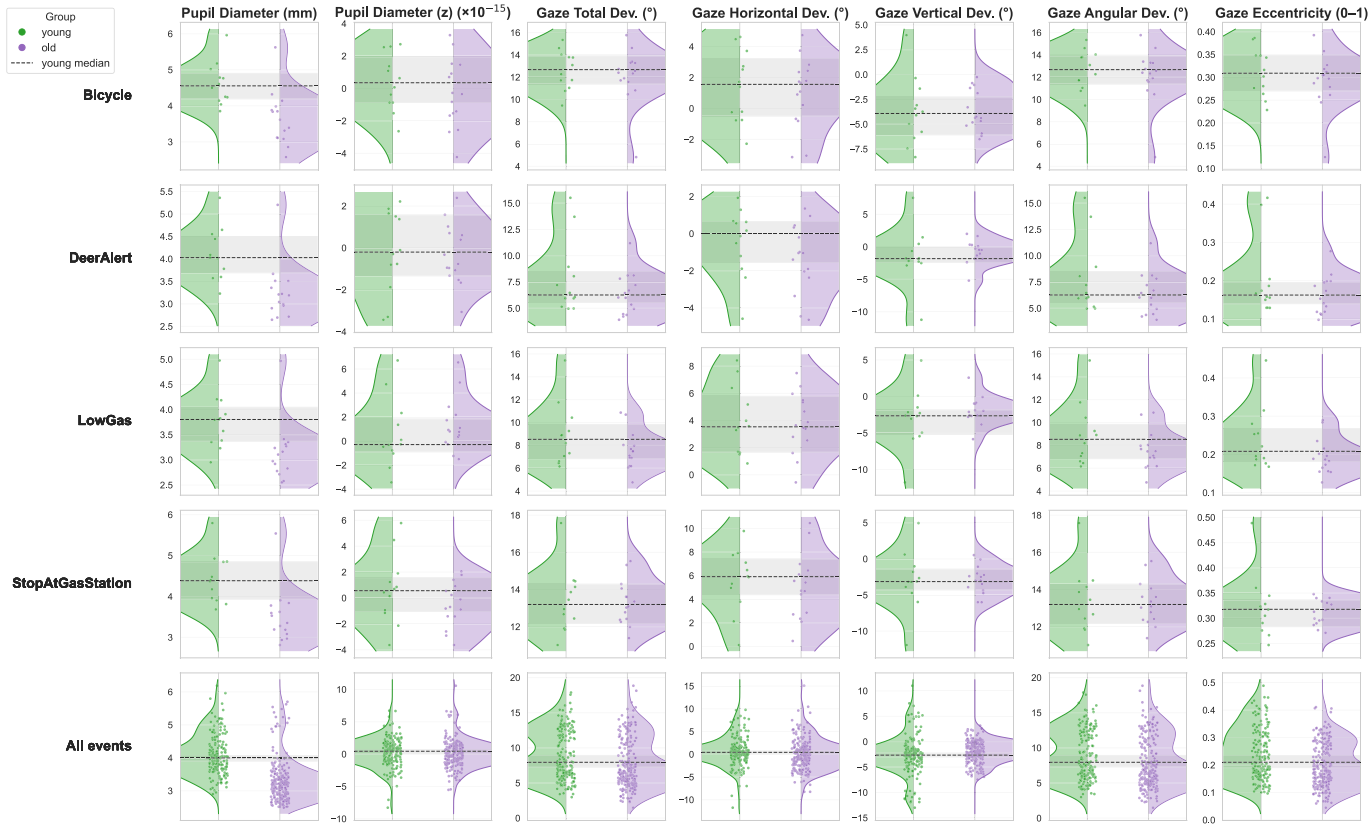


Fig. 2. Raincloud plots of physiological features across driving events by age group. Each subplot shows younger (green) and older (purple) participants with raw data points, kernel density estimates, and violin plots. Dashed black lines mark the medians of the ablation set (n = 20, non-HUD only) for reference. Features include pupil diameter (mm), normalized pupil diameter (*z*; scaled to a single power-of-ten as shown in the column title), horizontal, vertical, total, and angular gaze deviation (°), and gaze eccentricity (0–1; mean absolute gaze position relative to screen center).

Table 4
Age group comparisons using Mann–Whitney *U* tests with Holm correction. Results for the combined dataset ($n = 27$; 12 young, 15 old) are shown above, and for the ablation dataset ($n = 20$; 12 young, 8 old, non-HUD only) below.

Combined (27 participants)				
Feature	p (Holm)	Rank-Biserial	95% CI	Power
pupil_mean	< 0.001	−0.707	[−0.783, −0.632]	1.000
pupil_norm_mean	0.427	−0.141	[−0.307, 0.025]	0.684
gaze_position_mean	0.001	−0.238	[−0.341, −0.120]	0.997
gaze_angle_dev_mean	0.016	−0.197	[−0.307, −0.092]	0.974
Ablation (20 participants, non-HUD only)				
Feature	p (Holm)	Rank-Biserial	95% CI	Power
pupil_mean	< 0.001	−0.626	[−0.721, −0.532]	1.000
pupil_norm_mean	0.427	−0.141	[−0.296, 0.014]	0.684
gaze_position_mean	0.002	−0.259	[−0.366, −0.152]	0.993
gaze_angle_dev_mean	0.002	−0.243	[−0.369, −0.116]	0.974

deviation during `StopAtGasStation` ($p = 0.043$), where younger and older participants diverged in their gaze allocation strategies. Younger drivers showed more pronounced rightward scanning when approaching the gas station, while older drivers maintained more centered gaze patterns. No other event-by-age interactions survived multiple comparison correction in the combined dataset.

Within-group post-hoc analyses confirmed that younger participants ($n = 12$) showed multiple significant event contrasts in gaze metrics, including horizontal deviation between LowGas (median 3.44°) and StopAtGasStation (median 5.80°), as well as between monitoring and baseline driving ($p = 0.031$; Supplementary Table S2). Older participants showed no such within-group effects after correction, indicating more uniform scanning across events. When analyses were restricted to the four modeling events in the combined dataset, most younger-group contrasts diminished, with only the `StopAtGasStation` divergence remaining significant.

Younger participants consistently showed broader dispersion in horizontal gaze deviation across events, with typical ranges extending from approximately −10° to +15°, compared to −5° to +10° in older participants (Supplementary Table S3). These differences were statistically supported by per-event Mann–Whitney tests, which revealed significant group contrasts during `StopAtGasStation` and trend-level differences in other events. The descriptive ranges align with SHAP analyses (Section 4.2.4), where horizontal deviation variability emerged as the dominant feature driving classification. Together, these results indicate that younger drivers adopt more variable lateral scanning strategies, whereas older drivers maintain narrower, more centered gaze distributions.

4.2. Machine learning classification

4.2.1. Feature selection

A systematic feature ranking approach identified the most discriminative features for event classification using gradient boosted trees across leave-one-participant-out (LOPO) folds. From an initial set of 68 features computed from physiological signals, gaze-based metrics emerged as dominant predictors, with pupil-related features showing lower discriminative power.

The optimal feature subset size was determined by evaluating classification performance across multiple sizes ($K = 5, 7, 10, 12, 15$)

Table 5
XGBoost binary event vs. baseline classification performance (LOPO-CV, 27 participants). Metrics: Precision, Recall, F1, Balanced Accuracy, Support, and ROC-AUC. 95% CIs from 2000 bootstrap replicates at participant level. Event acronym: SAG = StopAtGasStation.

Event	Prec.	Rec.	F1	Bal. Acc.	Supp.
Bicycle	0.963	0.963	0.963 [0.933, 1.000]	0.963 [0.933, 1.000]	54
DeerAlert	0.708	0.630	0.667 [0.560, 0.769]	0.685 [0.594, 0.781]	54
LowGas	0.852	0.852	0.852 [0.786, 0.919]	0.852 [0.786, 0.921]	54
SAG	0.929	0.963	0.945 [0.909, 1.000]	0.944 [0.906, 1.000]	54
Mean	–	–	0.857 [0.820, 0.892]	0.861 [0.828, 0.896]	–
ROC-AUC (macro)			0.902 [0.873, 0.935]		–
ROC-AUC (micro)			0.921 [0.895, 0.951]		–

using aggregate gain ranking. Analysis identified $K = 7$ as optimal based on macro-F1 score maximization. This subset consisted entirely of gaze-based features: `gaze_angle_dev_auc`, `gaze_pos_mean`, `gaze_hdev_median`, `gaze_pos_median`, `gaze_pos_std`, `gaze_hdev_std`, and `gaze_hdev_auc`; formal definitions are provided in Supplementary Table S8.

To assess robustness, feature selection was conducted on both the full cohort ($n = 27$) and an ablation subset excluding HUD users ($n = 20$, non-HUD only). Both analyses independently identified the same seven features as optimal, confirming that HUD usage did not substantially alter selection outcomes while providing a homogeneous baseline for comparison.

These features capture horizontal deviation patterns (median, standard deviation, area under curve), gaze position characteristics (mean, median, standard deviation), and angular deviation dynamics. SHAP-based interpretation follows in Section 4.2.4.

4.2.2. Binary event detection

Binary classification on the combined dataset ($n = 27$) assessed how well individual events could be distinguished from baseline driving. Different events showed varying degrees of separability from baseline conditions (Table 5).

`StopAtGasStation` and `Bicycle` achieved the highest discrimination (F1 = 0.950), with `StopAtGasStation` showing particularly high ROC-AUC (0.986). `LowGas` showed moderate detection performance (F1 = 0.850, ROC-AUC = 0.948), while `DeerAlert` was most challenging (F1 = 0.737, ROC-AUC = 0.870). The lower performance for `DeerAlert` likely reflects its sudden onset and initial similarity to baseline forward driving.

Feature selection patterns varied by event: `LowGas` required 15 features for optimal discrimination, while `Bicycle` and `StopAtGasStation` achieved high performance with 12 features. Gaze position range and horizontal deviation metrics were consistently selected across events. These patterns align with event characteristics observed in statistical analyses, where horizontal deviation showed large differences between `StopAtGasStation` (Mdn = 5.85°) and baseline, while `LowGas` showed smaller shifts (Mdn = 3.53°).

4.2.3. Event classification performance

Multiclass discrimination between the four events achieved macro-F1 scores of 0.824 in the combined cohort and 0.850 in the ablation set using seven gaze features (Table 6). ROC-AUC values exceeded 0.95 in both cases, confirming event-level separability under LOPO evaluation.

Individual event performance varied systematically. `LowGas` achieved the highest F1-scores in both datasets (0.868 combined, 0.900 ablation), followed by `StopAtGasStation` (0.836 combined, 0.878 ablation) and `DeerAlert` (0.808 combined, 0.872 ablation). `Bicycle` consistently showed the lowest discrimination (0.786 combined, 0.750 ablation).

Table 6
XGBoost multiclass event classification performance (LOPO-CV, $K = 7$ gaze-based features). Results are reported for the combined 27-participant cohort and the 20-participant (no-HUD) ablation set. Metrics: Precision, Recall, F1-Score, Support, and ROC-AUC. 95% CIs (2000 bootstrap replicates at participant level) for macro-F1: Combined (27) [0.743, 0.898]; no-HUD (20) [0.751, 0.937].

	Combined (27)				No-HUD (20)			
Event	Prec.	Rec.	F1	Supp.	Prec.	Rec.	F1	Supp.
Bicycle	0.759	0.815	0.786	27	0.750	0.750	0.750	20
DeerAlert	0.840	0.778	0.808	27	0.895	0.850	0.872	20
LowGas	0.885	0.852	0.868	27	0.900	0.900	0.900	20
StopAtGas Station	0.821	0.852	0.836	27	0.857	0.900	0.878	20
Macro avg	0.826	0.824	0.824	108	0.850	0.850	0.850	80
Weighted avg	0.826	0.824	0.824	108	0.850	0.850	0.850	80
ROC-AUC (macro)	0.951 [0.914, 0.979]				0.964 [0.921, 0.993]			
ROC-AUC (micro)	0.954 [0.919, 0.980]				0.972 [0.934, 0.994]			

The ablation set showed higher performance across most events, with the difference most pronounced for `DeerAlert` (6.4% higher F1) and `Bicycle` (5.0% lower F1 compared to combined). These variations suggest that HUD inclusion introduced modest heterogeneity in gaze patterns, particularly for events involving lateral scanning or sudden responses. The consistency of selected features across both cohorts demonstrates that gaze dispersion metrics provide stable discrimination regardless of minor variations in sample composition.

4.2.4. SHAP analysis

SHAP (SHapley Additive exPlanations) values were computed to quantify feature contributions to event classification, calculating the marginal impact of each feature value on model predictions based on Shapley values from cooperative game theory.

Analysis of the seven gaze-based features revealed a clear contribution hierarchy (Fig. 3). Horizontal deviation standard deviation (`gaze_hdev_std`) emerged as the strongest feature (mean |SHAP| = 0.843), followed by horizontal deviation median (`gaze_hdev_median`, 0.646), horizontal deviation AUC (`gaze_hdev_auc`, 0.461), and gaze position standard deviation (`gaze_pos_std`, 0.351). The remaining features – angular deviation AUC (`gaze_angle_dev_auc`, 0.263), gaze position mean (`gaze_pos_mean`, 0.254), and gaze position median (`gaze_pos_median`, 0.216) – provided additional contributions (Supplementary Table S5). This ranking indicates that lateral scanning variability and central tendency were the dominant signals, with gaze eccentricity and angular dynamics contributing event-specific information.

Event-specific SHAP patterns illustrated how features differentiated driving scenarios (Fig. 4). For `StopAtGasStation`, horizontal deviation median values above 5° generated SHAP values exceeding 2.0, consistent with rightward scanning when approaching the station entrance. Horizontal deviation standard deviation contributed positively when elevated, reflecting variable scanning during visual search. `LowGas` showed the inverse pattern: horizontal deviation median values near or below 0° were associated with increased classification probability, whereas high values contributed negatively. These opposing patterns demonstrate how the same feature distinguished between rightward navigation scanning and central dashboard monitoring. Event-specific SHAP analyses were performed for all four events in the multiclass model. While `LowGas` and `StopAtGasStation` are shown here as representative cases, the corresponding SHAP results for the remaining two events (`Bicycle` and `DeerAlert`) are provided in the Supplementary Information (Figures S2 and S3).

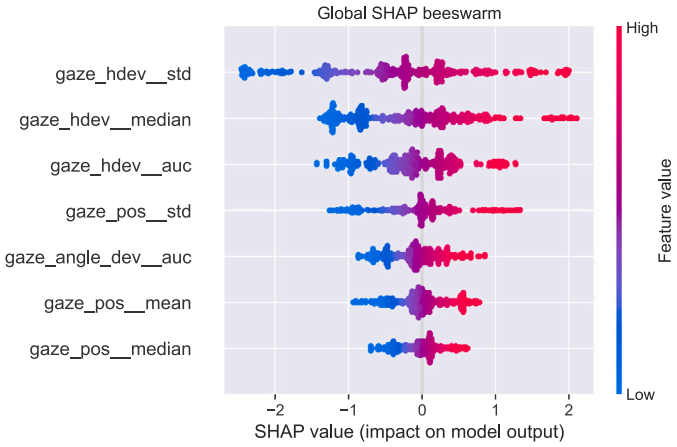


Fig. 3. Global SHAP values across all events in the combined cohort ($N = 27$). Features are ranked by mean absolute SHAP value. High feature values (red) increase the probability of certain events, while low values (blue) increase the probability of others. Horizontal deviation variability (`gaze_hdev_std`) and central tendency (`gaze_hdev_median`) showed the strongest contributions.

`DeerAlert` relied more strongly on angular gaze deviation AUC (`gaze_angle_dev_auc`) than other events, capturing sudden directional changes during the startle response. Gaze position variability (`gaze_pos_std`) also contributed, reflecting rapid visual search. `Bicycle` events were characterized by constrained gaze position measures, with low dispersion supporting classification (Supplementary Figure S2). `StopAtGasStation` showed the highest dependence on horizontal deviation median, while `LowGas` showed more balanced contributions across features (Supplementary Figure S3).

Age-stratified SHAP analysis revealed consistent feature hierarchies with different value ranges between groups (Supplementary Figures S4 and S5). Younger participants showed wider ranges for horizontal deviation standard deviation (`gaze_hdev_std`, approximately −10° to +15°) compared to older participants (−5° to +10°). This observation was statistically supported by per-event groupwise summaries (Supplementary Table S3), which confirmed broader horizontal deviation ranges in younger participants. Gaze position dispersion (`gaze_pos_std`) similarly spanned 0.10–0.40 in younger drivers versus 0.15–0.30 in older drivers. Despite these range differences, SHAP value magnitudes were comparable, indicating that the model mapped age-specific feature ranges to classification outputs.

The link between SHAP and behavioral data is further illustrated by age-specific response patterns. The combined dataset revealed a significant age × event interaction for horizontal gaze deviation during `StopAtGasStation` ($p = 0.043$), where younger participants exhibited more rightward scanning while older drivers maintained centered gaze. Within-group analyses showed that younger participants ($n = 12$) exhibited multiple significant contrasts in gaze metrics, including horizontal deviation between `LowGas` and `Wait` (Supplementary Table S2). No such contrasts were observed in older adults after correction, consistent with more uniform scanning across events.

4.3. Motif discovery

Time-series motif analysis revealed consistent temporal patterns across physiological signals in the combined dataset. Motif frequency remained stable (~0.12 bursts/s) across all conditions, with motifs spanning 92% of signal duration on average. This extensive coverage indicates that recurring patterns, rather than random fluctuations, dominated physiological behavior during driving.

4.3.1. Signal hierarchy and event modulation

Analysis of normalized motif amplitude revealed a consistent hierarchy across physiological signals. Gaze-derived motifs maintained robust amplitudes (3.0–3.4 z-scores) across all events, establishing gaze as a reliable carrier of event-related temporal structure. Pupil motifs showed comparable baseline strength (3.1–3.5 z-scores) but exhibited pronounced event-specific modulation.

This modulation was particularly evident during dashboard-focused events. During `LowGas` and `StopAtGasStation`, pupil motif amplitudes increased markedly in both age groups (young: 3.1–3.2 z-scores; old: 3.7–3.8 z-scores), matching or exceeding gaze motif levels. This selective enhancement during constrained visual tasks suggests that autonomic responses become more structured when gaze patterns are restricted by task demands, potentially compensating for reduced visual exploration through enhanced physiological organization.

4.3.2. Age-related differences in motif characteristics

Comparison between age groups revealed distinct yet complementary patterns in motif expression (Fig. 5). While both groups maintained robust gaze motif amplitudes (young: 3.0–3.3 z-scores, old: 3.0–3.4 z-scores), pupil motifs showed systematic age-related differences.

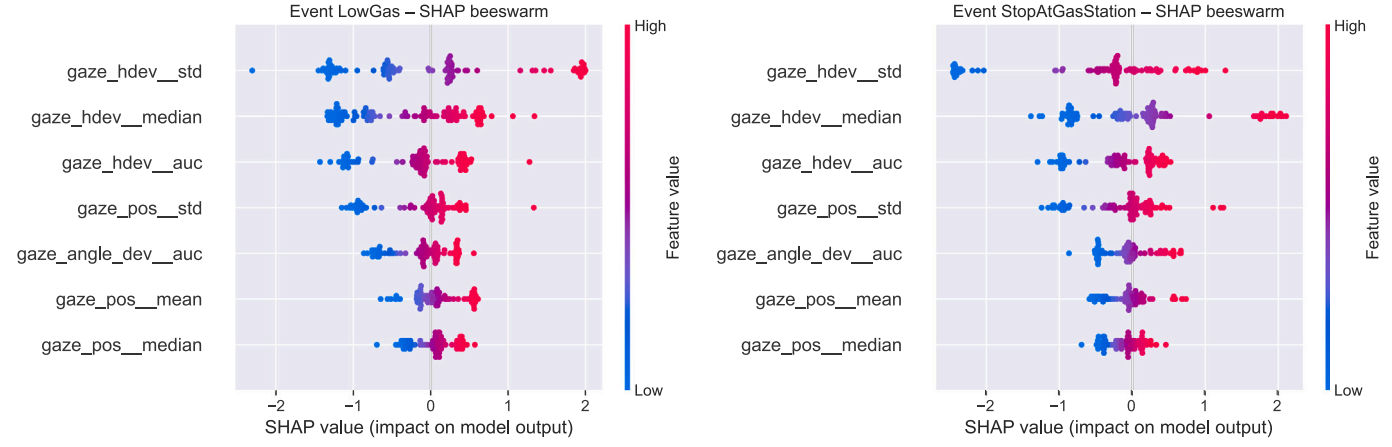


Fig. 4. Event-specific SHAP values for dashboard monitoring (`LowGas`, left) versus navigation (`StopAtGasStation`, right). For `LowGas`, low horizontal deviation values (blue) generate positive SHAP values, indicating central gaze. For `StopAtGasStation`, high horizontal deviation values (red) generate positive SHAP values, indicating rightward scanning.

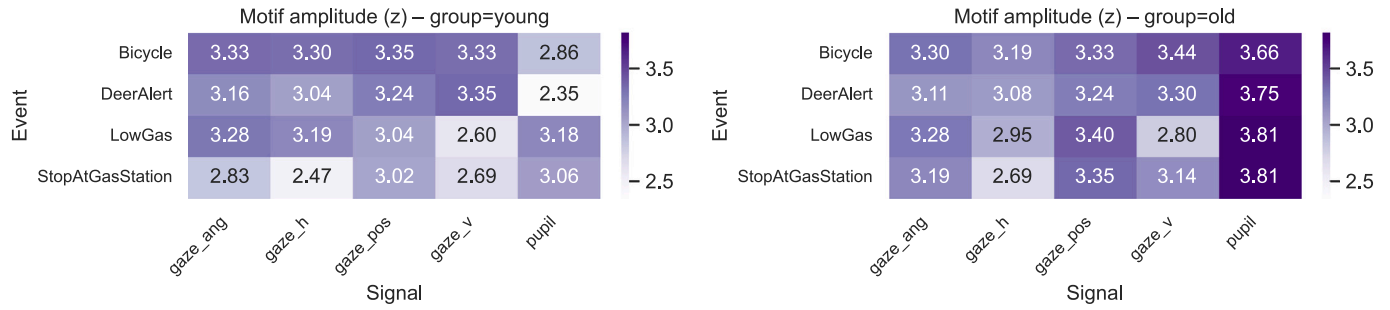


Fig. 5. Age-stratified motif amplitudes (z-normalized) with a standardized color scale across groups to permit direct comparison. Left: Younger drivers show uniformly strong gaze motifs across events, with pupil motifs generally weaker except for modest enhancement during `LowGas`. Right: Older drivers display elevated pupil motif amplitudes (notably during `Bicycle`, `LowGas`, and `StopAtGasStation`) in addition to robust gaze motifs.

The most pronounced difference emerged during `DeerAlert`, where older participants exhibited substantially stronger pupil motif amplitudes (3.75 z-scores) compared to younger participants (2.35 z-scores; $\Delta = 1.40$ z-scores). This pattern extended to dashboard monitoring events, with older drivers showing consistently elevated pupil motif amplitudes during `LowGas` (old: 3.81 [3.20–4.43] vs. young: 3.18 [2.70–3.72], $\Delta = 0.63$) and `StopAtGasStation` (old: 3.81 [3.42–4.24] vs. young: 3.06 [2.76–3.36], $\Delta = 0.75$).

Individual participant traces (Fig. 6) exemplified these group-level patterns. During `LowGas`, older participant 260 displayed remarkably stable pupil responses with structured oscillations, while younger participant 220 showed high-amplitude but less organized fluctuations. Gaze dynamics revealed complementary patterns: younger drivers exhibited variable horizontal and vertical excursions during `StopAtGasStation`, whereas older participants maintained more constrained scanning. These amplitude differences persisted throughout event duration, with no systematic drift between event halves (all $p_{\text{Holm}} > 0.40$).

4.3.3. Archetype analysis and temporal structure

To identify shared temporal patterns beyond individual motifs, archetypes were constructed through k-means clustering of pooled motif windows. These centroidal shapes captured collective temporal representations that generalized across participants within each event-signal combination.

Validation through intersection-over-union confirmed that archetypes accurately represented underlying motif structure. The mean IoU of 0.615 (median = 0.614) across all conditions increased to 0.710 with ±0.25 s temporal tolerance, demonstrating that archetypes captured the same temporal regions as individually-discovered motifs despite minor boundary variations (Supplementary Table S10).

Representative archetypes (Fig. 7) revealed event-specific temporal signatures. Horizontal gaze during `StopAtGasStation` showed consistent rightward scanning patterns. Angular gaze during `DeerAlert` and `Bicycle` captured characteristic steering corrections and hazard responses. Notably, `LowGas` pupil responses supported only four archetype clusters rather than five, suggesting more homogeneous autonomic patterns during sustained dashboard monitoring.

To relate these shapes to group-level consistency, each motif was assigned to its nearest archetype and a normalized membership entropy ($H_n \in [0, 1]$) was computed per group and panel (lower values indicate that motifs concentrate onto fewer archetypes). Archetype membership entropy (lower = more homogeneous) suggested that, during hazard events, older participants' motifs clustered more tightly to a few archetypes than younger participants (median $\Delta H_n \approx -0.16$), whereas groups were broadly similar for situational-awareness events (median $\Delta H_n \approx +0.02$); however, panel-wise permutation tests did not reach statistical significance after Benjamini–Hochberg FDR correction (Supplementary Table S11).

4.3.4. Statistical validation of age differences using prototypes

To statistically validate observed age differences, prototype-based comparisons were conducted using medoids—the actual observed windows that best represent each age group within event-signal panels. This approach preserves interpretability while enabling formal hypothesis testing.

Between-group separation, quantified as DTW distance between age-group medoids (Δ_{DTW}), revealed substantial differences across multiple event-signal combinations. The largest separations occurred for `Bicycle·gaze_pos` ($\Delta = 13.89$), `StopAtGasStation·pupil` (13.57), and `Bicycle·gaze_h` (11.48), with additional notable differences in `LowGas·gaze_ang` (11.44) and `StopAtGasStation·gaze_v` (11.40).

Global PERMANOVA testing confirmed age-related structure in the DTW space when controlling for event-signal composition (stratified: $R^2 \approx 1.3\%$, $p = 0.010$, $q = 0.020$), though this effect disappeared under free permutation ($p = 0.218$), indicating that age differences were event-specific rather than uniform. Panel-level analyses identified several event-signal combinations with nominal age effects ($p = 0.099$, $R^2 \approx 0.14$–0.18), though none survived multiple comparison correction ($q = 0.396$; Table 7).

Visual inspection of prototype shapes (Fig. 8) revealed the nature of these differences. Young and old medoids showed distinct temporal trajectories, with DTW barycenters (dashed lines) providing additional context while not affecting statistical inference. These prototype contrasts complement the archetype analysis by quantifying group differences while maintaining connection to actual observed patterns rather than synthetic centroids.

To situate these contrasts relative to previously reported archetype shapes (centroidal prototypes from k-means), note that the present analysis does not recluster motifs: it compares age-group prototypes within each event×signal panel and is therefore complementary to the curated archetype grid (which summarizes common shapes pooled across participants).

5. Discussion

This study investigated physiological responses to critical driving events through analysis of gaze dynamics and pupil responses during simulated driving. The convergent evidence from three analytical approaches demonstrates that gaze-based features can discriminate between multiple driving events with 82.4% multiclass accuracy (85.7% binary F1) under controlled conditions, while time-series motif discovery revealed recurring temporal patterns capturing both event-specific signatures and age-related physiological responses.

5.1. Gaze dynamics as primary event discriminators

The exclusive reliance on gaze features for optimal classification performance emerges from their behavioral relevance. Gaze patterns showed substantial event-driven modulation ($\eta^2 = 0.571$–0.582 for

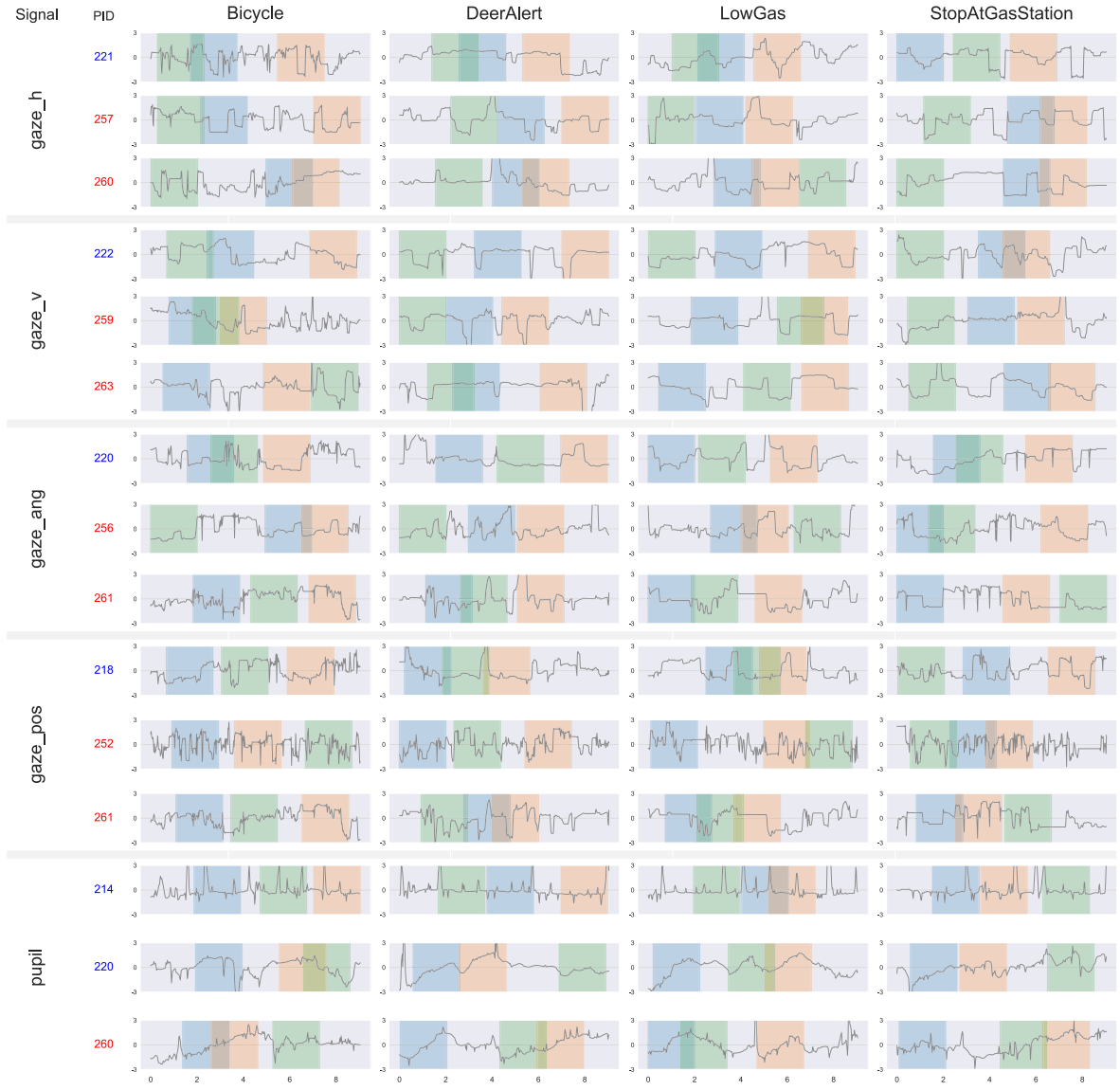


Fig. 6. Motif overlays for selected participants across events. Rows show individual participants (blue labels = older, red = younger), columns represent events. Each colored block indicates a detected motif burst in the z-normalized signal trace. Events differ in temporal gaze organization, with distinct motif patterns and variability profiles despite overlapping spatial demands. Selection highlights participants with clear age- or signal-specific contrasts, e.g., horizontal gaze shifts during `StopAtGasStation` and enhanced pupil motifs in older drivers during dashboard monitoring. Across participants, younger drivers (red PIDs) exhibited high-amplitude but irregular gaze excursions, whereas older drivers (blue PIDs) showed more constrained gaze alongside structured pupil oscillations, particularly during monitoring tasks.

range metrics), positioning gaze as the primary discriminative modality. This dominance aligns with Rahman et al. (2021), who achieved 92% accuracy distinguishing high from low cognitive load using 13 gaze features including fixation duration and saccade amplitude. However, the present work extends beyond binary load detection to discriminate between qualitatively different driving events, demonstrating that gaze patterns encode event-specific information beyond general workload levels.

The event-specific gaze patterns identified through SHAP analysis correspond directly to task demands: rightward scanning (median 5.85°) for gas station approach, leftward/central focus (median 3.53°) for dashboard monitoring during fuel warnings, and enhanced vertical deviation for deer hazard tracking. These patterns extend previous binary classification work by demonstrating multiclass event discrimination without requiring additional physiological sensors. The magnitude of gaze concentration during dashboard monitoring aligns with Recarte and Nunes (2003), who documented 23% reduction in horizontal scan variability during cognitive tasks, though the event-specific modulation observed here reveals that this narrowing varies systematically with task type rather than being a uniform response.

Notably, pupil diameter – despite validation as a cognitive load indicator – did not contribute to optimal classification when included even in the top-15 features. The z-normalization necessary for cross-participant comparisons likely removed individual baseline differences that make raw pupil diameter informative, suggesting a trade-off between population-level classification and individual monitoring. This contrasts with Čegovnik et al. (2018), who found pupil diameter increased monotonically from 3.2 mm to 4.1 mm with n-back difficulty, though they analyzed within-subject changes rather than cross-participant classification.

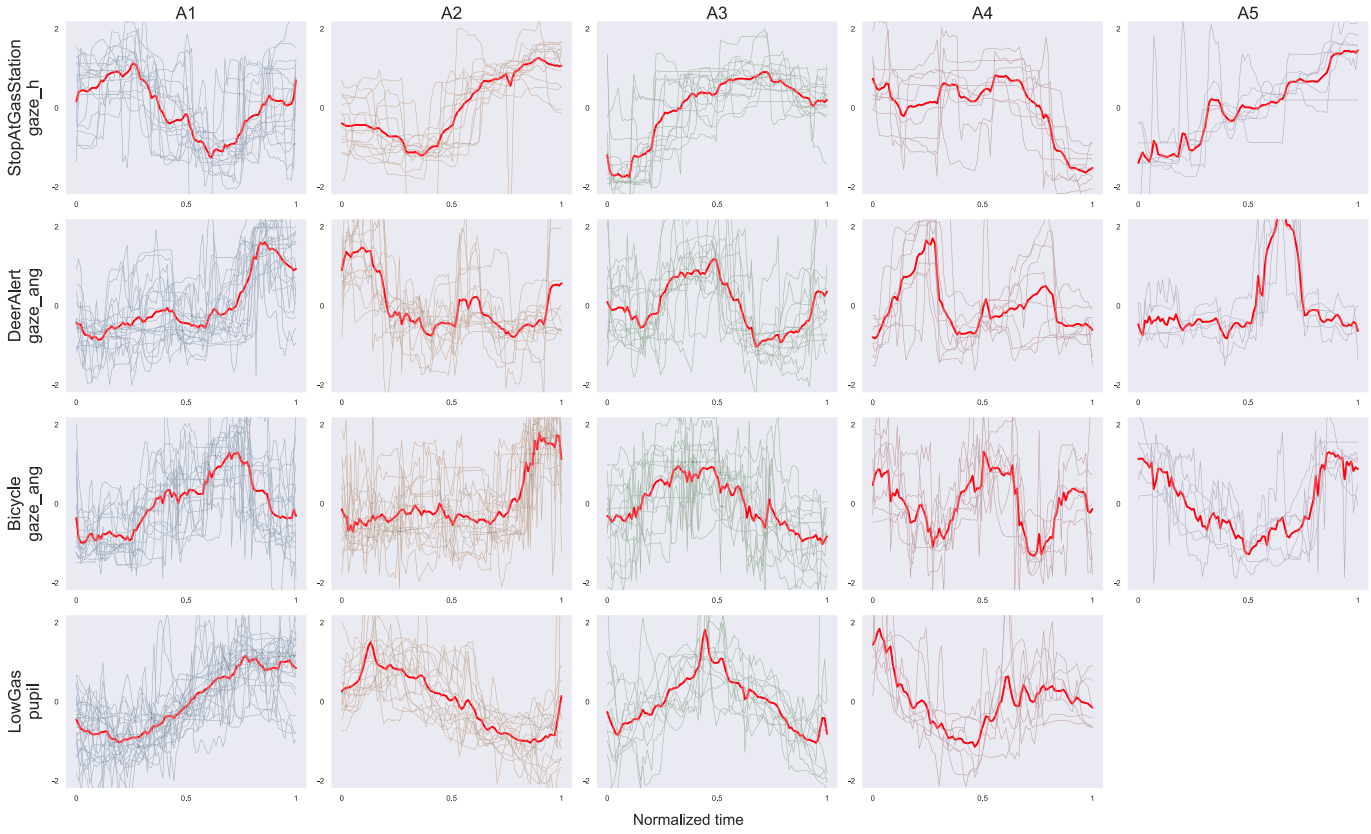


Fig. 7. Curated archetype grids across selected events and signal combinations. Archetypes capture consistent event-specific temporal structures across participants. Rows correspond to events (`StopAtGasStation · gaze_h`, `DeerAlert · gaze_ang`, `Bicycle · gaze_ang`, `LowGas · pupil`), columns to archetypes (A1–A5). Each panel shows the archetype centroid (red) with member motif spans in faint colors. Axes are normalized time (0–1) and z-normalized signal units. Archetypes reveal recurring temporal prototypes for monitoring, hazard response, steering, and physiological adaptation. Across rows, consistent event-specific structures emerge: lateral scanning shifts during navigation (row 1), angular corrections during hazard response (rows 2–3), and sustained pupil elevations during dashboard monitoring (row 4).

Table 7
PERMANOVA on the DTW motif space (young vs. old). Global tests use all windows; the stratified analysis permutes labels within event×signal. The lower block lists the per-panel analyses with the strongest nominal effects (none remain significant after BH–FDR across 20 panels).

Scope	N	R^2	p	q
Global (unstratified)	280	0.013	0.218	0.218
Global (stratified by event×signal)	280	0.013	0.010	0.020
Exemplary panels (per-panel PERMANOVA; nominal, not FDR-significant)				
Bicycle·gaze_pos (panel)	14	0.158	0.099	0.396
DeerAlert·gaze_ang (panel)	14	0.165	0.099	0.396
DeerAlert·gaze_pos (panel)	14	0.174	0.099	0.396
LowGas·pupil (panel)	14	0.182	0.099	0.396
StopAtGasStation·gaze_v (panel)	14	0.141	0.099	0.396

5.2. Temporal pattern discovery through motif analysis

From a behavioral research perspective, the temporal pattern analysis revealed that driver responses constitute structured behavioral strategies rather than random fluctuations, varying systematically across events and age groups. Motifs – recurring subsequences within individual signals – captured 92% of signal duration on average, indicating that drivers employ consistent physiological strategies such as repeated gaze shifts between road and dashboard or recurrent pupil dilations during monitoring tasks. The amplitude hierarchy (gaze motifs at 3.0–3.4 z-scores versus pupil motifs at 3.1–3.5 z-scores) independently validates the gaze dominance observed in classification while revealing organized temporal structure in both signals.

The construction of archetypes – canonical temporal patterns shared across participants – demonstrated collective behavioral prototypes that generalize beyond individual strategies. With 60–71% overlap between individually-discovered motifs and group-level archetypes, these patterns capture both commonality and diversity in driver responses. Specific archetypes revealed interpretable behaviors: lateral scanning shifts during `StopAtGasStation` navigation, angular deviations during `DeerAlert` responses, and slower pupil dilations during `LowGas` dashboard monitoring. Unlike traditional scalar features that discard temporal information, these preserved response shapes provide mechanistic insight into how driving events unfold physiologically over time.

This approach parallels Silva and Henriques (2021) TripMD system, which extracted driving motifs from vehicle telematics without labels. However, the physiological motifs captured here reveal driver-internal responses rather than vehicle dynamics, exposing the cognitive and attentional processes underlying maneuver execution. The successful construction of event-specific archetypes demonstrates that unsupervised methods can identify meaningful physiological signatures without prior labeling, suggesting potential for detecting novel or unexpected events in naturalistic driving.

5.3. Age-related physiological response patterns

The findings reveal distinct age-related physiological patterns in driving, though the limited sample (12 younger, 15 older participants)

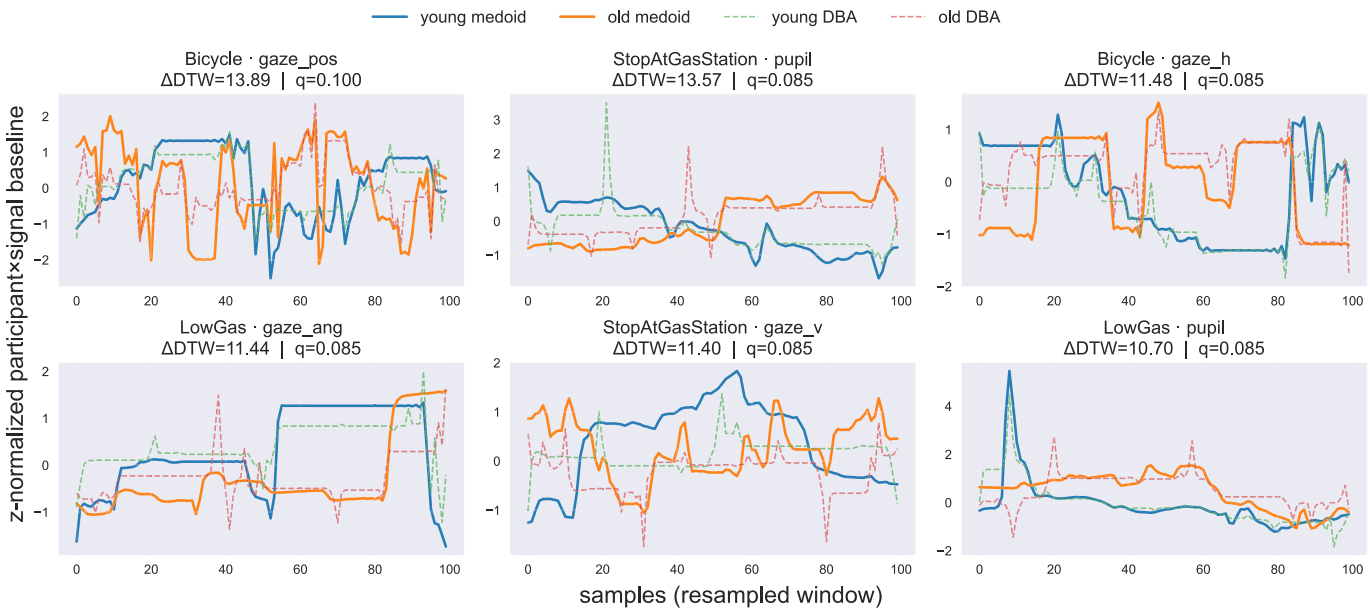


Fig. 8. Selected event×signal overlays of age-group prototypes, which highlight stable average gaze dynamics per event. Solid lines are group medoids (blue = young, orange = old); dashed, faint lines are DTW barycenters (DBAs) saved per panel. Y-axis: z-normalized signal (participant×signal baseline); X-axis: samples (resampled window, 0–100). Subplot titles report event×signal, along with ΔDTW and BH-FDR q values. Medoids drive all inference; DBAs are shown for shape context only.

requires cautious interpretation. While the study replicated known visual search patterns – including 28–35% reduced scan paths in older drivers (consistent with Maltz and Shinar, 1999) – the motif analysis revealed systematic age-related differences: elevated pupil motif amplitudes in older drivers during specific events.

These patterns align with previous research, which posits that older adults recruit additional resources to maintain performance (Reuter-Lorenz and Cappell, 2008). While our data cannot definitively establish compensation without performance measures, the observed physiological patterns are consistent with compensatory frameworks described in the cognitive aging literature. The elevated pupil organization coinciding with constrained gaze in older drivers represents an age-related adaptation pattern warranting further investigation with concurrent performance assessments.

The most pronounced age difference emerged during sustained attention tasks, with older drivers showing pupil motif amplitudes of 3.88 z-scores during `LowGas` compared to 3.18 in younger drivers ($\Delta = 0.70$). This pattern, observed across the older participant group, aligns with theoretical frameworks such as the CRUNCH hypothesis proposed by Reuter-Lorenz and Cappell (2008), which posits that older adults recruit additional neural resources to maintain performance. These findings extend this framework from central neural measures to peripheral autonomic responses, though whether this represents compensation requires performance data not collected in this study.

The co-occurrence of constrained gaze with enhanced pupil motifs reveals an intriguing pattern warranting further investigation. As documented by Underwood (2007) and confirmed in the present data, older drivers exhibited reduced gaze variance alongside elevated pupil motif amplitudes during dashboard-focused events (LowGas: 3.81 vs 3.18 z-scores; StopAtGasStation: 3.81 vs 3.06 z-scores). This pattern resonates with Kim et al. (2023)'s demonstration of larger task-evoked pupil dilations in older adults under high cognitive load, extending their laboratory findings to simulated driving tasks. Whether this co-occurrence represents a compensatory multi level physiological response requires concurrent performance assessment in larger samples.

The temporal dynamics observed align with Getzmann et al. (2018)'s EEG findings of increased frontal theta recruitment in older drivers, though manifested here through peripheral autonomic rather than central neural measures. This contrasts with Zahabi and Kaber (2018), who found no physiological compensation despite older drivers reporting higher subjective workload. The discrepancy likely stems from methodological differences: their aggregated measures would miss the temporal dynamics revealed by motif analysis.

Age-related conclusions in this study are grounded in direct statistical comparisons (Section 4.1.4) and motif-based temporal analyses (Section 4.3) rather than classifier performance, which does not isolate age effects when both age groups are pooled. To assess whether gaze-based features generalize across age groups, supplementary age-stratified analyses evaluated within-group classification (LOPO within each age group) and cross-group transfer (train on one group, test on other). Results showed comparable performance across conditions (within-group: old macro-F1 = 0.802, young = 0.731; cross-group: old→young = 0.816, young→old = 0.817; Supplementary Table S7), indicating that gaze-based features capture event-related patterns that transfer across age groups rather than encoding age-specific signatures that would limit generalizability.

5.4. Spatial arrangement and event classification: beyond location-based discrimination

Although the driving scenario used a fixed road geometry, event discrimination in the present study cannot be reduced to absolute gaze position or spatial layout. Importantly, the multiclass model predicts predefined event labels based on gaze response patterns and does not perform semantic object recognition; results reflect discrimination of event-associated physiological signatures rather than identification of scene entities such as gas stations or bicycles. The classifier does not encode predefined spatial regions or left–right mappings, but relies on within-participant normalized features capturing dynamic properties of gaze behavior, including deviation magnitude, variability, dispersion, and temporal integration. Importantly, several events share similar broad viewing demands (e.g., forward roadway monitoring or intermittent dashboard checks), yet remain discriminable. If classification were driven primarily by spatial positioning, such events would be expected to collapse, which is not observed. These findings indicate that the model distinguishes events based on how visual attention is organized

and evolves over time, rather than where participants look in absolute terms.

We elaborate this argument through multiple converging lines of evidence: feature interpretation, discrimination of spatially overlapping events, age-stratified performance consistency, temporal motif structure, and generalization considerations across different road layouts.

5.4.1. Multi-dimensional feature space beyond spatial location

The seven gaze-based features selected through systematic feature ranking capture temporal dynamics, variability, and integrated behavior rather than static spatial coordinates. Our optimal feature set consists of: `gaze_hdev_std` (horizontal deviation standard deviation), `gaze_hdev_median`, `gaze_hdev_auc` (area under curve), `gaze_pos_std` (position standard deviation), `gaze_angle_dev_auc`, `gaze_pos_mean`, and `gaze_pos_median`.

If classification depended solely on spatial location, central tendency measures (median, mean) would dominate feature importance. Instead, SHAP analysis (Fig. 3, Supplementary Table S5) reveals that `gaze_hdev_std` (|SHAP| = 0.843) emerges as the strongest discriminator, followed by `gaze_hdev_median` (0.646). This 30% difference in importance demonstrates that the model prioritizes how variably participants scan (captured by standard deviation) over where gaze centers (captured by median). The median provides complementary directional context—disambiguating whether high variability reflects leftward versus rightward scanning—but does not drive classification independently. Targeted ablation confirms this interpretation: removing all three central-tendency features (`gaze_hdev_median`, `gaze_pos_mean`, `gaze_pos_median`) caused only a modest performance drop (macro-F1: 0.824→0.806, $\Delta = 0.018$), whereas removing both variability features alone caused substantial degradation (macro-F1: 0.824→0.599, $\Delta = 0.225$; Supplementary Table S9).

The area-under-curve features (`gaze_hdev_auc`, `gaze_angle_dev_auc`) integrate gaze trajectory over time, capturing sustained directional biases and angular momentum of eye movements. These temporal accumulations distinguish between brief glances and prolonged monitoring episodes, dynamics orthogonal to instantaneous spatial position. Similarly, `gaze_pos_std` quantifies the spatial extent of visual exploration regardless of its directional bias, measuring whether attention remains focal or expands peripherally.

To assess the relative contributions of feature types, targeted ablations were conducted by removing feature subsets and retraining the model. Removing central-tendency features (`gaze_hdev_median`, `gaze_pos_mean`, `gaze_pos_median`) produced a modest performance drop (macro-F1: 0.824→0.806, $\Delta = 0.018$). Removing AUC features (`gaze_hdev_auc`, `gaze_angle_dev_auc`) yielded a comparable drop (0.824→0.798, $\Delta = 0.027$). However, removing both left only two standard deviation features and caused substantial degradation (0.824→0.599, $\Delta = 0.225$); this larger drop reflects both the loss of directional context and reduced feature dimensionality. Notably, removing central-tendency features (3 features) caused a smaller drop than removing AUC features (2 features), despite removing more features, suggesting that AUC provides modestly more unique information. Under fixed-length windows, AUC and central-tendency measures are strongly collinear, both capturing distributional bias, which explains their partial substitutability. The critical finding is that variability—how broadly participants scan—dominates classification (SHAP: `gaze_hdev_std` = 0.843), while directional features provide necessary but partially substitutable context for interpreting that variability across events (Supplementary Table S9).

5.4.2. Event-specific behavioral signatures transcend spatial overlap

The discrimination between events sharing similar spatial characteristics provides direct evidence against pure location-based classification. Both LowGas and StopAtGasStation involve rightward and dashboard attention, yet achieve distinct multiclass classification performance (F1 = 0.868 and 0.836, respectively). Statistical analyses reveal contrasting gaze patterns: LowGas exhibits constrained central monitoring (horizontal deviation median = 3.53°, lower variability), whereas StopAtGasStation involves active rightward visual search (median = 5.85°, higher variability). The classifier discriminates between these monitoring versus search strategies despite shared spatial regions, confirming it captures qualitative differences in attentional allocation rather than mere directional preference.

Similarly, both DeerAlert and Bicycle events involve forward-directed hazards, yet classification distinguishes them through response dynamics rather than spatial quadrant. DeerAlert shows enhanced `gaze_angle_dev_auc` reflecting sudden directional shifts characteristic of startle responses, combined with increased vertical deviation as participants track the crossing trajectory. In contrast, Bicycle events are characterized by constrained horizontal gaze variability and focused forward attention, consistent with sustained monitoring of a predictable moving hazard. Compared to DeerAlert, Bicycle events show smoother gaze trajectories and reduced dispersion, reflecting anticipatory tracking rather than reactive reorientation.

5.4.3. Age-stratified patterns reveal strategy, not location

Age-related differences in classification performance and gaze patterns further demonstrate that the model captures relative behavioral strategies rather than absolute spatial coordinates. Older participants exhibited uniformly constrained gaze across all events (Supplementary Table S3: horizontal range approximately −5° to +10°) compared to younger participants (−10° to +15°), reflecting well-documented reductions in visual scanning (Maltz and Shinar, 1999; Underwood, 2007). If classification relied primarily on absolute spatial location, age groups with substantially different scanning ranges should show divergent classification accuracies and the model would be calibrated to one group's spatial distribution. Instead, both groups achieved comparable multiclass performance: combined dataset (82.4% accuracy) and ablation subset (85.0%). This consistency indicates the model captures relative scanning strategies, including the extent of exploration, variability around individual baselines, and directional biases scaled to each participant's typical behavior.

Critically, age groups also exhibited complementary physiological patterns despite constrained gaze in older drivers. Older participants demonstrated elevated pupil motif amplitudes during dashboard-focused events (LowGas: 3.88 versus 3.18 z-scores in younger drivers, $\Delta = 0.70$; StopAtGasStation: 3.81 versus 3.06, $\Delta = 0.75$). This enhancement of autonomic organization during periods of spatially constrained gaze represents a distinct physiological signature independent of eye position, reflecting structured temporal patterns in pupillary responses that correlate with task demands rather than gaze direction.

5.4.4. Motif analysis validates temporal structure beyond position

The time-series motif analysis provides independent validation that physiological responses contain event-specific temporal structure orthogonal to spatial location. Matrix profile-based motif discovery identified recurring patterns spanning 92% of signal duration. Gaze motif amplitudes remained robust across events (3.0–3.4 z-scores), while pupil motifs showed selective enhancement during monitoring tasks (3.1–3.8 z-scores), with particularly pronounced increases in older drivers during dashboard-focused events. The temporal shapes of these motifs, visualized through archetypes (Fig. 7), reveal structured response dynamics distinct from spatial positioning. LowGas pupil motifs exhibit gradual drifts and sustained elevated states characteristic of prolonged cognitive monitoring, contrasting sharply with the rapid deflections and recovery patterns observed during DeerAlert startle responses. These temporal dynamics encode how physiological systems organize over time in response to event demands, information that cannot be reduced to "looking left versus right."

5.4.5. Generalization considerations across road layouts

We acknowledge that our fixed event sequence, with StopAtGasStation consistently appearing on the right and LowGas requiring leftward dashboard glances, raises questions about generalizability to different road layouts. However, several aspects of our methodology support cross-context transferability. Dispersion-based gaze metrics (standard deviation, range) and integrated measures (area under curve) capture the extent and variability of scanning behavior independent of absolute direction. A rightward gas station search in right-side driving manifests as leftward scanning in left-side contexts, but the underlying feature signature (elevated horizontal standard deviation, increased angular deviation) remains equivalent. The model learns to recognize "broad lateral exploration" versus "constrained central monitoring" as abstract patterns that are expected to transfer across mirrored layouts.

Additionally, all features undergo z-normalization relative to each participant's baseline (participant × signal), ensuring that feature values reflect deviations from individual typical behavior rather than absolute screen coordinates. This normalization approach explicitly accounts for inter-individual differences in scanning habits, baseline gaze positions, and sensor calibration offsets. A model trained on right-biased layouts encodes "participants scan more rightward than their baseline" rather than "participants look at screen coordinate X," facilitating generalization to contexts where the same relative deviation occurs in different absolute directions. Despite these theoretical arguments, empirical validation remains necessary. Future work should evaluate classification performance across mirrored road layouts (left-side driving), randomized event positions within scenarios, and naturalistic driving where event locations cannot be controlled. Such validation would distinguish feature contributions, establishing which aspects of our gaze signatures reflect event-invariant attentional strategies versus context-specific spatial adaptations.

5.4.6. Synthesis: multi-dimensional physiological strategies

Converging evidence from feature interpretation, event-pair discrimination, age-stratified performance, and temporal pattern analysis demonstrates that our classification approach captures multi-dimensional physiological strategies rather than serving as a spatial location detector. The feature space encodes attentional breadth through the spatial extent of visual exploration (standard deviation, range), scanning dynamics through temporal integration of gaze trajectories (area under curve), directional biases through central tendencies modulated by task demands (median, mean), and autonomic modulation through pupillary response patterns independent of gaze position (motif amplitudes).

Events are discriminated through the combination of these dimensions: `LowGas` features constrained spatial exploration with enhanced pupil organization, `StopAtGasStation` combines broad lateral scanning with moderate pupil responses, and `DeerAlert` shows rapid angular shifts with selective vertical deviations. Spatial location serves as one component within this multi-dimensional space, but classification performance depends on capturing the full constellation of physiological and behavioral adaptations that characterize each driving event. This interpretation aligns with theoretical frameworks emphasizing that gaze behavior reflects underlying cognitive strategies (Yarbus, 1967; Underwood, 2007), and extends this perspective by demonstrating that machine learning models trained on gaze dynamics can extract event-specific signatures that generalize across individuals and age groups.

5.5. Behavioral implications: temporal patterns and adaptation strategies

The temporal analysis reveals patterns that may complement existing age-related driving literature. While studies like Cantin et al. (2009) and Yamani et al. (2016) document performance changes in older drivers (longer reaction times, missed secondary glances, centralized gaze), the motif analysis suggests these behavioral patterns co-occur with structured physiological responses that warrant further investigation.

The observed temporal dynamics of elevated pupil motif amplitudes during constrained gaze periods align with theoretical models of aging and performance. As Anstey et al. (2005) argued, constraints in visual attention and executive control combine with sensory decline to alter information processing strategies. The motif analysis provides one operationalization of these altered strategies, showing how they manifest as specific temporal patterns. However, whether these patterns represent compensation, adaptation, or simply age-related changes requires larger samples and performance measures not collected in this study.

These patterns resonate with findings from intervention studies. Romoser (2013)'s successful training of older drivers demonstrates behavioral plasticity, while Thorslund et al. (2019)'s documentation of compensatory behaviors in hearing-impaired older drivers suggests multiple adaptation pathways. The physiological findings may represent another facet of this adaptation, though this interpretation remains speculative given the study constraints.

5.6. Methodological contributions and convergent validation

The multi-method approach provided complementary insights essential for understanding complex driver behavior. Statistical analyses established signal reliability and event-wise differences. Machine learning achieved 82.4% multiclass accuracy distinguishing four driving events, with binary classification reaching 85.7% F1. Most critically, motif discovery revealed temporal organization invisible to aggregate measures, while prototype-based DTW comparisons enabled statistical validation of age differences using actual observed patterns.

From a behavioral research perspective, this methodological convergence captures not just what drivers do but how they organize responses over time. Traditional approaches that reduce temporal signals to scalar features miss the structured strategies drivers employ. By preserving temporal form through motifs and archetypes, the analysis revealed patterns that may inform understanding of driver behavior across different conditions and age groups.

The consistency between datasets (combined $n = 27$ and ablation $n = 20$), with HUD users showing no systematic differences, validates the robustness of event-specific signatures. The multiclass performance compares favorably to Hogervorst et al. (2014), though they performed binary workload classification rather than multi-event discrimination. The 9-second window reflects the temporal structure of driving events and physiological feature extraction requirements. Driving events differ fundamentally: some demand continuous scanning over extended periods (e.g., searching for a gas station while monitoring traffic), others involve focused, time-limited responses (e.g., reacting to a fuel warning). The selected duration accommodates extended patterns while maintaining event specificity and enables reliable computation of temporally aggregated gaze features (area under curve, standard deviation) and delayed pupil responses (typically 2–5 s post-onset). While Meteier et al. (2021) found 2–4 minute windows optimal for workload level classification across sustained conditions, discriminating qualitatively distinct driving events with varying temporal demands benefits from windows capturing complete behavioral sequences without aggregating across event types. Future work should systematically evaluate how window length interacts with event temporal structure to optimize response completeness versus temporal precision. This evaluation would establish optimal durations for multiclass event discrimination in naturalistic driving contexts.

5.7. Limitations and future directions

Several limitations warrant consideration. First, while our sample size ($n = 27$) aligns with standard driving simulator studies, larger samples would strengthen age-specific pattern detection. Second, the absence of driving performance metrics (e.g., lane-keeping variability, reaction times, speed maintenance) prevents definitive conclusions about

whether observed physiological patterns represent successful compensation or simply age-related variation. Third, the simulated environment, while controlled, differs from the unpredictability of naturalistic driving. Fourth, the 9-second window requirement excluded shorter safety-critical events that may be equally important for comprehensive driver monitoring.

Reliability analysis (Section 4.1.2) indicated that 94% of pupil observations and 87–92% of gaze observations exceeded predefined measurement-precision thresholds. The Tobii Pro Glasses 2 system with prescription lens inserts and individual calibration provided robust tracking in our synchronized motion-base simulator, as evidenced by high signal coverage and consistency across motion conditions. However, wearable eye-trackers can experience reduced accuracy during motion, with field studies documenting systematic effects of G-forces and vibrations on tracking precision in moving vehicles (Calvi et al., 2023; Sasalovici et al., 2025). While no participants reported motion sickness, subclinical vestibular effects cannot be ruled out; Lopes et al. (2020) demonstrated that cybersickness in VR contexts affects pupil position and blink rate, suggesting even subclinical disturbances could modulate oculomotor behavior.

Additionally, while no participants reported motion sickness symptoms, subclinical vestibular effects on gaze dynamics cannot be ruled out. Lopes et al. (2020) demonstrated that cybersickness in VR contexts affects pupil position and blink rate, suggesting that even subclinical vestibular disturbances could modulate oculomotor behavior. Future naturalistic driving studies should employ systematic tracking quality assessments to quantify performance under real-world motion conditions.

The inclusion of seven HUD users among older participants was addressed through sensitivity analysis, which revealed no significant display-related effects. This null finding suggests that our physiological signatures are robust to minor display variations, though future work should systematically investigate display modality effects.

The controlled environment contrasts with naturalistic studies like SHRP 2 (Shi et al., 2022), where real-world variability challenged even high-accuracy models. Future work should validate these temporal patterns in naturalistic settings with larger age-stratified samples and performance metrics. The unsupervised archetype approach could potentially discover novel event signatures in real-world data. Additionally, the misaligned depth camera limited body movement analysis to steering proximity measures rather than intended postural assessments.

5.8. Implications for driver monitoring systems

The findings suggest multiple pathways for driver monitoring systems. The multiclass discrimination performance using gaze features alone indicates potential for simplified sensor configurations, while temporal patterns provide richer behavioral understanding than aggregate features. Future systems could leverage motifs to detect strategy shifts or archetypes to identify departures from typical response patterns.

The potential age-related patterns observed – if validated in larger studies – suggest monitoring systems might benefit from considering different physiological response profiles across age groups. Systems could track whether the elevated pupil organization observed in older drivers maintains consistent levels, though determining optimal thresholds would require performance data not collected here. The motif-based approach offers specific operational advantages: motifs provide interpretable temporal signatures that could trigger context-specific alerts (e.g., detecting the onset of dashboard-focused monitoring versus active visual search), while archetypes establish normative response templates against which individual deviations could be quantified in real time. Unlike aggregate features that collapse temporal structure, motif metrics preserve the sequential organization of responses, potentially enabling earlier detection of strategy shifts before they manifest in degraded performance. The interpretable nature of motifs and archetypes could enable transparent systems where drivers understand detected patterns, though practical implementation requires further development.

5.9. Conclusions

This study demonstrates that gaze-based features achieve 82.4% accuracy in discriminating between four driving events in controlled simulator conditions, with binary classification reaching higher performance. The discovery of recurring motif patterns reveals structured temporal organization in driver responses. Age-stratified analyses uncovered intriguing patterns – including elevated pupil motif amplitudes in older drivers during constrained gaze periods – that warrant investigation in larger samples with performance measures. Despite limitations in sample size and experimental constraints, the convergence of statistical, machine learning, and temporal pattern discovery approaches provides both methodological innovation and preliminary insights for developing driver monitoring systems that leverage temporal dynamics alongside traditional features.

CRediT authorship contribution statement

Gregor Strle: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Conceptualization. **Kristina Stojmenova Pečečnik:** Writing – review & editing, Project administration, Funding acquisition, Data curation, Conceptualization. **Jaka Sodnik:** Writing – review & editing, Supervision, Project administration, Funding acquisition, Data curation, Conceptualization.

Ethics approval and consent to participate

The experiment was designed and conducted in accordance with the Code of Ethics for Researchers and the Guidelines for Ethical Conduct in Research with Human Subjects of the University of Ljubljana. Informed consent was obtained from all participants. Participation was voluntary, and participants were informed that they could withdraw from the experiment at any time without providing a reason.

Funding

This research was supported by the Slovenian Research Agency program ICT4QoL — Information and Communications Technologies for Quality of Life (P2-0246) and the Horizon Europe research and innovation program for the project FRODDO, grant agreement no. 101147819.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at doi: 10.1016/j.ijhcs.2026.103827.

Data availability

An anonymized package containing the data and code required to reproduce the reported analyses is available via OSF (view-only link: https://osf.io/5ckxg/overview?view_only=1faa44662d4b480688c2be0cddea7487). Representative video recording of the four analyzed events, captured from the driver's perspective using the Tobii Pro Glasses 2 scene camera (1920×1080 at 25 fps), is also provided. The repository contains the analysis scripts, processed data, and supplementary materials necessary to reproduce the reported results.

References

Anstey, K.J., Wood, J., Lord, S., Walker, J.G., 2005. Cognitive, sensory and physical factors enabling driving safety in older adults. Clin. Psychol. Rev. 25, 45–65. https://doi.org/10.1016/j.cpr.2004.07.008. https://www.sciencedirect.com/science/article/pii/S0272735804001345.

Biernacki, M.P., Lewkowicz, R., 2021. How do older drivers perceive visual information under increasing cognitive load? Significance of personality on-road safety. Accid. Anal. Prev. 157, 106186. https://doi.org/10.1016/j.aap.2021.106186

Calvi, A., D'Amico, F., Vennarucci, A., 2023. Comparing eye-tracking system effectiveness in field and driving simulator studies. The Open Transportation Journal 17, e187444782301191. https://doi.org/10.2174/18744478-v17-e230404-2022-49

Cantin, V., Lavallière, M., Simoneau, M., Teasdale, N., 2009. Mental workload when driving in a simulator: effects of age and driving complexity. Accid. Anal. Prev. 41, 763–771. https://doi.org/10.1016/j.aap.2009.03.019. https://www.sciencedirect.com/science/article/pii/S0001457509000633.

Charness, N., 2008. Aging and human performance. Hum. Factors 50, 548–555. https://doi.org/10.1518/001872008X312161. pMID: 18689066.

Chen, T., Guestrin, C., 2016. XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Association for Computing Machinery, New York, NY, USA, pp. 785–794, https://doi.org/10.1145/2939672.2939785

Clark, H., Feng, J., 2017. Age differences in the takeover of vehicle control and engagement in non-driving-related activities in simulated driving with conditional automation. Accid. Anal. Prev. 106, 468–479. https://doi.org/10.1016/j.aap.2016.08.027. https://www.sciencedirect.com/science/article/pii/S000145751630313X.

Cooper, J.M., Wheatley, C.L., McCarty, M.M., Motzkus, C.J., Lopes, C.L., Erickson, G.G., Baucom, B.R.W., Horrey, W.J., Strayer, D.L., 2020. Age-related differences in the cognitive, visual, and temporal demands of in-vehicle information systems. Front. Psychol. 11, 1154. https://doi.org/10.3389/fpsyg.2020.01154

Derafshi, Z., Krejtz, I., Duchowski, A.T., 2024. Impact of cognitive impairment on driving behaviour and route choices of older drivers. Sci. Rep. 14, https://doi.org/10.1038/s41598-024-63663-y

Doshi, A., Trivedi, M.M., 2009. On the roles of eye gaze and head dynamics in predicting driver's intent to change lanes. IEEE Trans. Intell. Transp. Syst. 10, 453–462. https://doi.org/10.1109/TITS.2009.2026673

Duchowski, A.T., Krejtz, K., Gehrer, N.A., Bafna, T., Bækgaard, P., 2020. The low/high index of pupillary activity. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, Association for Computing Machinery, New York, NY, USA, pp. 1–12, https://doi.org/10.1145/3313831.3376394

Getzmann, S., Arnau, S., Karthaus, M., Reiser, J.E., Wascher, E., 2018. Age-related differences in pro-active driving behavior revealed by EEG measures. Front. Hum. Neurosci. 12, 321. https://doi.org/10.3389/fnhum.2018.00321

González-Ortega, D., González-Díaz, J., Díaz-Pernas, F.J., Martínez-Zarzuela, M., Antón-Rodríguez, M., 2017. 3D kinect-based gaze region estimation in a driving simulator. In: Proceedings of the 11th International Conference on Ubiquitous Computing and Ambient Intelligence (UCAmI 2017), pp. 789–795, https://doi.org/10.1007/978-3-319-67585-5_76

Guo, F., Ai, Y., Qu, S., 2025. Intersection challenges for older drivers: the impact of aging on visual cognition and driving efficiency at crossroads. Traffic Injury Prevention 0 1–10. https://doi.org/10.1080/15389588.2025.2463615. pMID: 40036651.

Harris, C.R., Millman, K.J., van der Walt, S.J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N.J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M.H., Brett, M., Haldane, A., del Río, J.F., Wiebe, M., Peterson, P., Gérard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C., Oliphant, T.E., 2020. Array programming with NumPy. Nature 585, 357–362. https://doi.org/10.1038/s41586-020-2649-2

He, D., Wang, Z., Khalil, E.B., Donmez, B., Qiao, G., Kumar, S., 2022. Classification of driver cognitive load: exploring the benefits of fusing eye-tracking and physiological measures. Transp. Res. Rec. 2676, 670–681. https://doi.org/10.1177/03611981221090937

Hess, E.H., Polt, J.M., 1964. Pupil size in relation to mental activity during simple problem-solving. Science 143, 1190–1192. https://doi.org/10.1126/science.143.3611.1190

Hogervorst, M.A., Brouwer, A.M., Van Erp, J.B., 2014. Combining and comparing EEG, peripheral physiology and eye-related measures for the assessment of mental workload. Front. Neurosci. 8, 322.

Hooge, I.T.C., Niehorster, D.C., Hessels, R.S., Benjamins, J.S., Nyström, M., 2023. How robust are wearable eye trackers to slow and fast head and body movements? Behav. Res. Methods 55, 4128–4142. https://doi.org/10.3758/s13428-022-01935-1

Hu, X., Lodewijks, G., 2021. Exploration of the effects of task-related fatigue on eye-motion features and its value in improving driver fatigue-related technology. Transp. Res. F Traffic Psychol. Behav. 80, 150–171. https://doi.org/10.1016/j.trf.2021.03.014. https://www.sciencedirect.com/science/article/pii/S1369847821000681.

Hunter, J.D., 2007. Matplotlib: a 2D graphics environment. Computing in Science & Engineering 9, 90–95. https://doi.org/10.1109/MCSE.2007.55

IntelRealSense, 2020. Intel realsense d400 series depth modules, vision processors, and depth cameras product datasheet. Depth accuracy specification https://www.intelrealsense.com/ (June 2020). Rev. approximately ± 2 %.

Kahneman, D., 1973. Attention and Effort. Prentice-Hall, Englewood Cliffs, NJ.

Kahneman, D., Beatty, J., 1966. Pupil diameter and load on memory. Science 154, 1583–1585. https://doi.org/10.1126/science.154.3756.1583

Kim, J., Lee, J., Jun, S.B., Sung, J.E., 2023. Pupillometry as a window to detect cognitive aging in the brain. Biomed. Eng. Lett. 14, 91–101. https://doi.org/10.1007/s13534-023-00315-6

Koay, H.V., Chuah, J.H., Chow, C.O., Chang, Y.L., 2022. Detecting and recognizing driver distraction through various data modalities using machine learning: a review, recent advances, simplified framework and open challenges (2014–2021). Eng. Appl. Artif. Intell. 115, 105309. https://doi.org/10.1016/j.engappai.2022.105309

Kraljevska, M., Hlaváčková-Schindler, K., Miklautz, L., Plant, C., 2025. Motif discovery framework for psychiatric EEG data classification. arXiv preprint arXiv:2501.04441.

Law, S.M., 2019. STUMPY: a powerful and scalable Python library for time series data mining. The Journal of Open Source Software 4, 1504.

Lopes, P., Tian, N., Boulic, R., 2020. Eye thought you were sick! Exploring eye behaviors for cybersickness detection in VR. In: Proceedings of the 13th ACM SIGGRAPH Conference on Motion, Interaction and Games, Association for Computing Machinery, New York, NY, USA, https://doi.org/10.1145/3424636.3426906

Lundberg, S.M., Lee, S.I., 2017. A unified approach to interpreting model predictions. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. Curran Associates Inc, Red Hook, NY, USA, pp. 4768–4777.

Maltz, M., Shinar, D., 1999. Eye movements of younger and older drivers. Hum. Factors 41, 15–25. https://doi.org/10.1518/001872099779577282

McKinney, W., 2010. Data structures for statistical computing in Python. In: van der Walt, STÉFAN., Millman, JARROD. (Eds.), Proceedings of the 9th Python in Science Conference, pp. 56–61, https://doi.org/10.25080/Majora-92bf1922-00a

Mehler, B., Reimer, B., Pohlmeyer, A., Coughlin, J., 2008. The Association Between Heart Rate Reactivity and Driving Performance Under Dual Task Demands in Late Middle Age Drivers. J.F.

Mehler, B., Reimer, B., Coughlin, J.F., Dusek, J.A., 2009. Impact of incremental increases in cognitive workload on physiological arousal and performance in young adult drivers. Transp. Res. Rec. 2138, 6–12. https://doi.org/10.3141/2138-02

Mehler, B., Reimer, B., Coughlin, J.F., 2012. Sensitivity of physiological measures for detecting systematic variations in cognitive demand from a working memory task: an on-road study across three age groups. Hum. Factors 54, 396–412. https://doi.org/10.1177/0018720812442086. pMID: 22768642.

Merat, N., de Waard, D., 2014. Human factors implications of vehicle automation: current understanding and future directions. Transp. Res. F Traffic Psychol. Behav. 27, 193–195. https://doi.org/10.1016/j.trf.2014.06.004

Meteier, Q., Capallera, M., Ruffieux, S., Angelini, L., Abou Khaled, O., Mugellini, E., 2021. Classification of drivers' workload using physiological signals in conditional automation. Front. Psychol. 12, 596038. https://doi.org/10.3389/fpsyg.2021.596038

Michelaraki, E., Katrakazas, C., Kaiser, S., Brijs, T., Yannis, G., 2023. Real-time monitoring of driver distraction: state-of-the-art and future insights. Accid. Anal. Prev. 192, 107241. https://doi.org/10.1016/j.aap.2023.107241

Niehorster, D.C., Hessels, R.S., Benjamins, J.S., Nyström, M., Hooge, I.T.C., 2024. Glassesvalidator: a data quality tool for eye tracking glasses. Behav. Res. Methods 56, 1476–1484. https://doi.org/10.3758/s13428-023-02105-5

Nilsson, E.J., Aust, M.L., Engström, J., Svanberg, B., Lindén, P., 2018. Effects of cognitive load on response time in an unexpected lead vehicle braking scenario and the detection response task (drt). Transp. Res. F Traffic Psychol. Behav. 59, 463–474. https://doi.org/10.1016/j.trf.2018.09.026. https://www.sciencedirect.com/science/article/pii/S1369847817306198.

Rahman, H., Ahmed, M.U., Barua, S., Funk, P., Begum, S., 2021. Vision-based driver's cognitive load classification considering eye movement using machine learning and deep learning. Sensors 21, 8019. https://doi.org/10.3390/s21238019

Rauh, A., Petermeijer, S.M., Heine, J., Bünger, C., Hoedemaeker, D., 2020. Multimodal driver fatigue detection: eye movements, electrooculography and heart rate. Transp. Res. F Traffic Psychol. Behav. 69, 149–162. https://doi.org/10.1016/j.trf.2020.04.008

Recarte, M.A., Nunes, L.M., 2003. Mental workload while driving: effects on visual search, discrimination, and decision making. J. Exp. Psychol. Appl. 9, 119–137. https://doi.org/10.1037/1076-898X.9.2.119

Ren, Y., Li, Y., Loftus, T.J., 2024. Identifying acute illness phenotypes via deep temporal interpolation and clustering network on physiologic signatures. Sci. Rep. 14, https://doi.org/10.1038/s41598-024-59047-x

Reuter-Lorenz, P.A., Cappell, K.A., 2008. Neurocognitive aging and the compensation hypothesis. Current Directions in Psychological Science 17, 177–182. https://doi.org/10.1111/j.1467-8721.2008.00570.x

Romoser, M.R.E., 2013. The long-term effects of active training strategies on improving older drivers' scanning in intersections: a two-year follow-up to romoser and fisher (2009). Hum. Factors 55, 278–284. https://doi.org/10.1177/0018720812457566. pMID: 23691824.

Sasalovici, M., Zeqiri, A., Schramm, R.C., Ariza Nunez, O.J., Jansen, P., Freiwald, J.P., Colley, M., Winkler, C., Rukzio, E., 2025. Bumpy ride? Understanding the effects of external forces on spatial interactions in moving vehicles. In: Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems, Association for Computing Machinery, New York, NY, USA, https://doi.org/10.1145/3706598.3714077

Sharma, P.K., Chakraborty, P., 2024. A review of driver gaze estimation and application in gaze behavior understanding. Eng. Appl. Artif. Intell. 133, 108117. https://doi.org/10.1016/j.engappai.2024.108117. https://www.sciencedirect.com/science/article/pii/S0952197624002756.

Shi, L., Qian, C., Guo, F., 2022. Real-time driving risk assessment using deep learning with XGBoost. Accid. Anal. Prev. 178, 106836. https://doi.org/10.1016/j.aap.2022.106836. https://www.sciencedirect.com/science/article/pii/S0001457522002718.

Silva, M.I., Henriques, R., 2021. TripMD: driving patterns investigation via motif analysis. Expert Syst. Appl. 184, 115527. https://doi.org/10.1016/j.eswa.2021.115527

Solovey, E.T., Zec, M., Garcia Perez, E.A., Reimer, B., Mehler, B., 2014. Classifying driver workload using physiological and driving performance data: two field studies. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Association for Computing Machinery, New York, NY, USA, pp. 4057–4066, https://doi.org/10.1145/2556288.2557068

Sun, L., Liu, Z., Sun, M.T., 2015. Real-time gaze estimation with a consumer depth camera. Inf. Sci. 294, 18–31. https://doi.org/10.1016/j.ins.2015.05.027

Tao, X., Gao, D., Zhang, W., 2024. A multimodal physiological dataset for driving behaviour analysis. Sci. Data 11, https://doi.org/10.1038/s41597-024-03222-2

Thorslund, B., Nygårdhs, S., Malicka, A., Black, A., Hickson, L., Wood, J., 2019. Exploring older adults hearing and vision and driving – the Swedish study. Transp. Res. F Traffic Psychol. Behav. 64, 274–284. https://doi.org/10.1016/j.trf.2019.04.011

Tobii, 2020. Pro glasses 2 eye tracker. https://www.tobii.com/products/discontinued/tobii-pro-glasses-2 (Online 19 May 2023).

Underwood, G., 2007. Visual attention and the transition from novice to advanced driver. Ergonomics 50, 1235–1249. https://doi.org/10.1080/00140130701318707

Vallat, R., 2018. Pingouin: statistics in Python. J. Open Source Softw. 3, 1026. https://doi.org/10.21105/joss.01026

Čegovnik, T., Stojmenova, K., Jakus, G., Sodnik, J., 2018. An analysis of the suitability of a low-cost eye tracker for assessing the cognitive load of drivers. Appl. Ergon. 68, 1–11. https://doi.org/10.1016/j.apergo.2017.10.011. https://www.sciencedirect.com/science/article/pii/S0003687017302326.

Virtanen, P., 2019. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat. Methods 17, 261–272. https://doi.org/10.1038/s41592-019-0686-2

Vortmann, L., Krenz, F., Böhmer, M., Schmidt, A., 2021. Imaging human attention: learning from gaze time-series represented as images. Front. Neurosci. 15, 711290. https://doi.org/10.3389/fnins.2021.711290

Walker, B.N., Stanton, N.A., Carter, H., 2019. Viability of combining gaze and physiological measures to assess driver trust in level-2 automation. Transp. Res. F Traffic Psychol. Behav. 62, 1714–1722. https://doi.org/10.1016/j.trf.2019.01.017

Wang, Y., Reimer, B., Dobres, J., Mehler, B., 2014. The sensitivity of different methodologies for characterizing drivers' gaze concentration under increased cognitive demand. Transp. Res. F Traffic Psychol. Behav. 26, 227–237. https://doi.org/10.1016/j.trf.2014.08.003. https://www.sciencedirect.com/science/article/pii/S136984781400120X.

Wang, A., Huang, C., Wang, J., He, D., 2024. The association between physiological and eye-tracking metrics and cognitive load in drivers: a meta-analysis. Transp. Res. F Traffic Psychol. Behav. 104, 474–487. https://doi.org/10.1016/j.trf.2024.06.014. https://www.sciencedirect.com/science/article/pii/S1369847824001475.

Waskom, M.L., 2021. seaborn: statistical data visualization. J. Open Source Softw. 6, 3021. https://doi.org/10.21105/joss.03021

Wei, W., Xue, Q., Yang, X., Du, H., Wang, Y., Tang, Q., 2023. Assessing the cognitive load arising from in-vehicle infotainment systems using pupil diameter. In: Cross-Cultural Design: 15th International Conference, CCD 2023, Held as Part of the 25th International Conference, HCII 2023, Copenhagen, Denmark, July 23–28, 2023, Proceedings, Part II. Springer-Verlag, Berlin, Heidelberg., pp. 440–450, https://doi.org/10.1007/978-3-031-35939-2_33

Yamani, Y., Samuel, S., Roman Gerardino, L., Fisher, D.L., 2016. Extending analysis of older drivers' scanning patterns at intersections. Transp. Res. Rec. 2602, 10–15. https://doi.org/10.3141/2602-02

Yarbus, A.L., 1967. Eye Movements and Vision. Plenum Press, New York.

Zahabi, M., Kaber, D.B., 2018. Effect of police mobile computer terminal interface design on Officer driving distraction. Appl. Ergon. 67, 26–38. https://doi.org/10.1016/j.apergo.2017.09.006

Zargari Marandi, R., Madeleine, P., Omland, Ø., 2018. Eye movement characteristics reflected fatigue development in both young and elderly individuals. Sci. Rep. 8, https://doi.org/10.1038/s41598-018-31577-1

Zhang, N., Yang, C., Fard, M., 2022. Measured increases in steering entropy may predict when performance will degrade: a driving simulator study. Transp. Res. F Traffic Psychol. Behav. 91, 87–94. https://doi.org/10.1016/j.trf.2022.10.006. https://www.sciencedirect.com/science/article/pii/S136984782200225X.

Zhang, R., Tan, Z., Lin, Z., Zhang, R., Liu, C., 2025. Exploring the trust and behavior of experienced advanced driver assistance system drivers: an on-road study. Accid. Anal. Prev. 217, 108071. https://doi.org/10.1016/j.aap.2025.108071. https://www.sciencedirect.com/science/article/pii/S0001457525001575.

Zhou, F., Yang, X.J., de Winter, J.C.F., 2022. Using eye-tracking data to predict situation awareness in real time during takeover transitions in conditionally automated driving. IEEE Trans. Intell. Transp. Syst. 23, 2284–2295. https://doi.org/10.1109/TITS.2021.3069776