Multimodal Physiological Fusion for Driver Fatigue Detection: Evidence from a 160-Hour Naturalistic Truck Study

Zhijie Yi[1], Chenglin Liu[1], Xiang Chang[2], Boyu Jin[3], Dengbo He[1,3,*]

***Abstract*— Reliable fatigue detection in long-haul commercial truck operations is critical for traffic safety. Yet current driver monitoring systems struggle with inter-driver variability and cumulative physiological changes during extended driving hours. In this work, we conducted a naturalistic data collection in long-haul trucks. In total, 160 hours of physiological and subjective Karolinska Sleepiness Scale (KSS) data from eight professional drivers were collected, with approximately 20 hours per driver, to investigate how multimodal fusion of physiological data may address inter-subject variability in driver state monitoring. Under a Leave-One-Subject-Out (LOSO) protocol, we show that integrating electrocardiogram (ECG) and respiration (RESP) improves detection reliability across individuals. ECG+RESP fusion achieved a Precision-Recall Area Under the Curve of 0.249, a 25.5% relative improvement over the best single-modality baseline (0.199). It is possible that fusing the respiration signal with the cardiovascular modality may filter out the subject-specific false alarms from a single modality. These findings suggest that multimodal physiological features are needed for long-haul driver fatigue monitoring, potentially supporting more reliable fatigue detection in truck operations.**

I. Introduction

Long-haul goods transportation by heavy-duty trucks is the backbone of modern logistics, yet it faces a high risk of crashes due to prolonged driving and irregular rest. Specifically, unlike short-distance commuting, long-haul operations involve continuous driving sessions lasting more than 10 hours, often occurring during circadian troughs [1], [2]. Thus, driver fatigue is especially prevalent in long-haul truck operations. Given the catastrophic outcome of driver fatigue, a robust driver monitoring system (DMS) becomes especially important [3].

Previous research has tried to estimate drivers' fatigue based on four categories of measures [4], including driving-performance-based measures, such as steering wheel inputs and lane keeping performance [5], [6], subjective measures (e.g., Karolinska Sleepiness Scale (KSS) [7], [8]), eye-tracking measures (e.g., PERCLOS, [9], [10]), and physiological measures. Given the advantage of physiological measures over other measures (e.g., more sensitive to low levels of fatigue as compared to performance measures [11], less sensitivity to eye-tracking measures as compared to eye-tracking measures [4] and better capability to provide continuous measures as compared to subjective measures), the potential early fatigue detection capabilities [12], and the possibility of non-intrusive physiological signal collection technologies [13], various physio-based DMS systems targeting driver fatigue have been proposed.

Although previous research has proposed algorithms to estimate driver fatigue [14], [15], [12], a gap remains between laboratory-based and real-world findings. Most previous studies were based on driving simulator experiments or controlled field studies with short continuous exposure, often <2 hours [16], [14], [15]. For example, Ingre et al. [10] examined KSS, blink duration, and lane deviation in a 2-hour high-fidelity driving simulator experiment. Similarly, Anund et al. [17] collected KSS, blink, and driving performance measures during 2-hour controlled motorway sessions, repeated across daytime and nighttime conditions. While these settings provide experimental control, they under-represent long-haul operations where fatigue accumulates gradually over many hours [18], [19]. Thus, it is necessary to build and validate driver fatigue monitoring algorithms using a real-world dataset.

At the same time, high inter-subject variability remains a major hurdle to the generalizability of the DMS algorithms. For example, a model that performs well on a specific driver may fail when encountering an individual with a different baseline [20], [21]. This is especially true for algorithms based on physiological signals, where individual heterogeneity can dominate. Thus, in this work, we present an empirical study using a self-collected 160-hour naturalistic truck dataset involving eight professional drivers. Driver fatigue detection algorithms based on physiological measures were proposed and validated. To ensure our results reflect deployment-oriented generalization, we avoided random window-level splits and instead employed a Leave-One-Subject-Out (LOSO) protocol. Our main contributions are as follows:

- We present a 160-hour physiological dataset captured during real-world freight operations, reflecting the real-world complexity of sparse fatigue-related events at the window level (11.9% prevalence).
- Multimodal ECG+RESP fusion enhances detection robustness across individuals, achieving a 25.5% relative improvement in Precision-Recall Area Under the Curve (PR-AUC) over single-modality baselines.

II. Related Work

A. Fatigue Monitoring and Multimodal Fusion of Physiological Measures

The detection of driver fatigue via physiological signals has been extensively explored. Single-modality approaches rely heavily on the electrocardiogram (ECG) to extract heart rate variability (HRV) [22], [12], or on the electroencephalogram

[1]Thrust of Intelligent Transportation, Systems Hub, The Hong Kong University of Science and Technology (Guangzhou), Guangzhou, China
[2]Cornell Tech, Cornell University, New York, USA
[3]Thrust of Robotics and Autonomous Systems, Systems Hub, The Hong Kong University of Science and Technology (Guangzhou), Guangzhou, China
*Corresponding author: dengbohe@hkust-gz.edu.cn.

(EEG) to capture brainwave transitions. Respiratory features (RESP) have also been introduced to assess physical lethargy [23]. Recognizing the limitations of individual sensors, recent studies have shifted toward multimodal fusion. However, the majority of these fusion frameworks were validated under highly controlled conditions [24]. The efficacy of multimodal synergy in commercial truck operations in naturalistic settings remains under-investigated in the literature.

B. Limitations of Short-Duration and Simulator-Based Studies

A critical limitation in the current fatigue detection literature is the reliance on driving simulations or short-duration on-road tests [16], [14]. In these settings, drivers typically experience "acute fatigue" induced by deliberate sleep deprivation, and the continuous driving task rarely exceeds two hours. While such protocols offer high experimental control, they fundamentally misrepresent the "cumulative fatigue" experienced by professional long-haul drivers [18], [19]. During real-world 10-hour shifts, physiological signals undergo gradual baseline shifts, such as macroscopic alterations in cardiovascular tone, due to seated posture, physical exhaustion, and circadian rhythms [2]. Thus, algorithms developed under short-duration protocols may overfit to acute sleepiness patterns and generalize poorly to long-haul freight operations, where time-on-task and circadian effects evolve over time [16], [12], [20], [21].

C. Cross-Subjects Evaluation Protocols

As fatigue monitoring systems transition from academic prototypes to industrial deployments, reliable cross-subjects performance has emerged as a primary bottleneck [20], [21]. Many existing studies report exceptionally high detection accuracies, such as Area under the Receiver Operating Characteristic Curve (ROC-AUC) $>$ 0.9, by utilizing random window-level data splitting. This evaluation protocol inadvertently allows data from the same driver to appear in both the training and test sets, thereby leaking individual physiological identities and temporal autocorrelations to the classifier. Consequently, the reported performance is often an optimistic upper bound rather than a reflection of deployment readiness. To address this flaw, it is necessary to adopt a LOSO cross-validation protocol to evaluate multimodal fusion across different drivers.

III. Methodology

A. Participants

Eight professional long-haul heavy-duty truck drivers participated in this study. They were all male, aged from 37 to 47 years old ($M = 42.6$, $SD = 3.9$). Drivers were recruited from an active freight operation and had extensive experience with extended-duration highway driving and schedule-based work routines. They were also required to self-report being healthy, without acute illness, cardiovascular conditions, or medications known to influence autonomic activity. Each driver completed two long-haul sessions (outbound and return), providing repeated, long-duration exposure to investigate within-session changes in alertness.

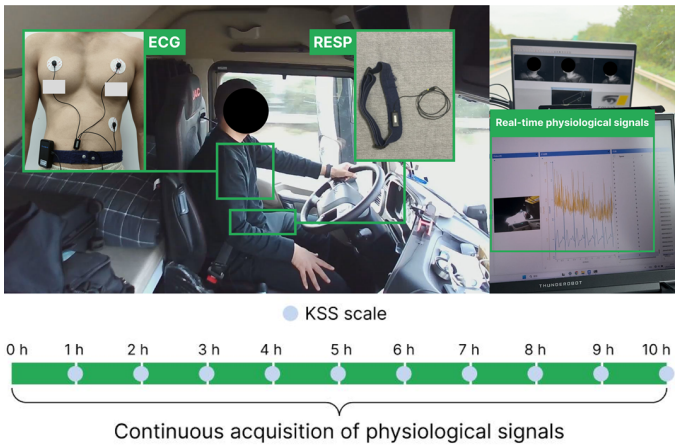


Fig. 1: Experiment design and data acquisition process.

Data collection was governed by a university-industry agreement specifying privacy protections and permitted analyses. Participation was voluntary, and data were de-identified. The study was approved by the Human and Artifacts Research Ethics Committee at the Hong Kong University of Science and Technology (Guangzhou) (No. HKUST(GZ)-HSP-2023-0013).

B. Driving Route and Experimental Truck

Data were collected on an operational route between Wuxi, Jiangsu, and Qingdao, Shandong (Fig. 2). The outbound journey ($\sim$750 km) averaged 10.7 h ($SD = 0.57$), while the return journey, including a stopover in Weifang ($\sim$795 km), averaged 11.8 h ($SD = 0.48$). To standardize supervision, time zero was defined as the expressway entry, ensuring that feature extraction and labeling occur during the first $\sim$10 hours of highway driving per session.

A JAC K7 five-axle heavy-duty truck tractor equipped with the PlusAI pilot-assisted driving system and a standard cargo container was used in the study (Fig. 3). The truck is approximately 18 m long and weighs 35 t. The PlusAI system was developed by Suzhou PlusAI Technology Co., Ltd. and designed for highway use. Consistent with Society of Automotive Engineers (SAE, [25]), the system can be categorized as Level 2, a driver-support system: when the ADAS was activated, the human driver remained responsible for supervising the roadway and taking over when needed. It should be noted that, as the ADAS engagement behavior was not strictly manipulated and more and more trucks are now equipped with SAE Level-2 ADAS, we did not differentiate ADAS effects on driver fatigue and did not include ADAS status as an input feature. Thus, the results of this study should be interpreted as fatigue monitoring in an L2-equipped long-haul truck in general, without differentiating ADAS engagement.

C. Data Acquisition and Subjective Assessment

Physiological signals were recorded continuously. ECG and respiration (RESP) were sampled at 200 Hz using the PhysioLab by Ergoneers, Germany, and Human Research Tool

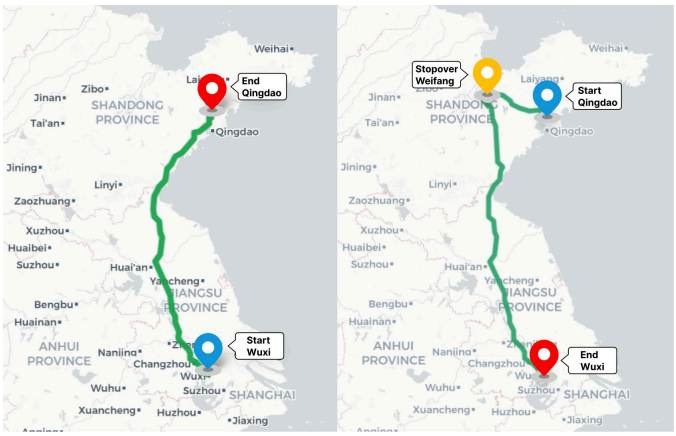


Fig. 2: Driving routes. Left: outbound route from Wuxi to Qingdao. Right: return route from Qingdao to Wuxi with a stopover at Weifang.



Fig. 3: The JAC K7 heavy-duty truck used in the study.

software by InfoTech, China, under a shared acquisition clock, ensuring alignment before feature extraction. A trained researcher in the co-driver seat monitored signal quality throughout each session. It should be noted that the cardiovascular and respiratory features were adopted here, as they have been found to be associated with driver fatigue in previous studies [26]. More importantly, recent advancements have made remote monitoring of cardiovascular and respiratory signals possible [13]. Thus, the cardiovascular and respiratory measures have real-world applications.

Concurrent subjective alertness was assessed hourly via brief verbal prompts using the KSS, ranging from 1 to 9 [27]. Prompting began after highway entry, hourly for up to 10 hours, and drivers were asked to evaluate their overall sleepiness during the preceding 60 minutes.

D. Data Processing

Raw physiological signals were preprocessed using the NeuroKit2 package [28]. For ECG data, a band-pass filter was applied to eliminate baseline wander, high-frequency power-line noise, and muscle artifacts. R-peaks were then identified using NeuroKit2's QRS detection algorithms to derive heart rate (HR) and heart rate variability (HRV) metrics.

Continuous RESP signals were band-pass filtered (0.1–0.5 Hz) to isolate the respiratory frequency band and attenuate high-vibration truck-cabin noise. The cleaned waveforms were processed to detect inspiratory peaks and expiratory troughs,

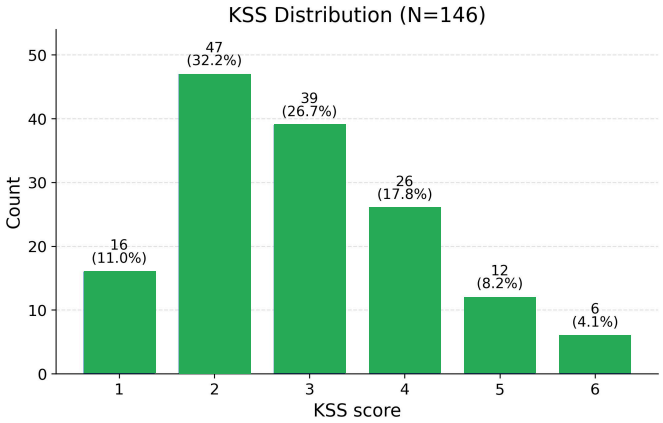


Fig. 4: Distribution of hourly KSS across all drive sessions ($N = 146$).

enabling calculation of instantaneous respiratory rate and breath amplitude before window-level aggregation.

E. Label Mapping and Dataset Partition

While KSS was collected hourly, physiological features were extracted over 5-minute windows. To preserve temporal consistency, each hourly KSS score was mapped backward to the twelve preceding 5-minute windows. Unlike nearest-neighbor assignments, this mapping prevents temporal data leakage by ensuring that models do not use future subjective states to label past physiology.

In this work, "fatigue" was operationally defined as KSS-based reduced alertness, rather than clinically verified sleep onset, microsleep, or crash risk. We used KSS ≥ 5 as an early-warning label because KSS 5 marks the transition between alertness and sleepiness. Further, the observed KSS ranges fell within mild to moderate (maximum KSS = 6), and a stricter threshold would leave too few positive events (i.e., reduced alertness) for cross-driver learning. Thus, this threshold should be interpreted as a sensitive reduced-alertness threshold rather than a severe-fatigue threshold.

Under the 10-hour horizon, 146 hourly KSS reports were collected (91.3% of the theoretical maximum 160), of which 12.3% were labeled as reduced alertness (KSS ≥ 5; Fig. 4). After label mapping and removal of windows with signal dropouts or poor quality, the final analyzable dataset contained $N = 1,665$ labeled 5-minute windows. The final window-level positive samples accounted for 11.9% of the dataset, which also served as the chance-level baseline for PR-AUC [29].

F. Feature Extraction and Pre-Processing

From the continuous data streams, we extracted HR, RMSSD (root mean square of successive differences between adjacent normal-to-normal R-R intervals), LF/HF (low-frequency-to-high-frequency ratio), respiratory rate, and respiratory amplitude. To mitigate between-driver scaling differences and time-on-task physiological baseline drifts, features (x_t) were normalized using their first-hour mean ($\bar{x}_{\mathrm{Hour1}}$)

as $r_t = x_t/\bar{x}_{\text{Hour1}}$. Right-skewed HRV features (RMSSD and LF/HF) were corrected using a log transformation: $z_t = \log(x_t/\bar{x}_{\text{Hour1}})$.

To evaluate multimodal integration, we designed three combinations of features:

1) **ECG only**: HR, RMSSD, and LF/HF.
2) **RESP only**: respiratory rate and respiration amplitude.
3) **ECG + RESP**: HR, RMSSD, LF/HF, respiratory rate, and respiration amplitude.

G. Classifiers and Cost-Sensitive Learning

We evaluated four classical classifiers: Support Vector Machine (i.e., SVM, with RBF kernel), K-Nearest Neighbors (KNN, $k = 7$), Random Forest (RF), and XGBoost. To handle the data imbalance at the 5-minute window level (11.9% prevalence), we employed Cost-Sensitive Learning for the SVM, RF, and XGBoost [30], [31]. Specifically, dynamic class weighting scaled the misclassification penalty for the minority fatigue class proportionally to its scarcity in each training fold, prioritizing sparse fatigue detection without hallucinating data. Note that the Cost-Sensitive Learning was not applied to KNN, as the loss function was not used.

H. Cross-Drivers Evaluation Protocol (LOSO-CV) for Model Validation

We simulated real-world fleet deployments by employing a Leave-One-Subject-Out Cross-Validation (LOSO-CV) protocol [32]. In each of the 8 folds, all windows from one driver were held out as the test set, while the remaining 7 drivers formed the training set. All data pre-processing was performed exclusively on the training set to prevent data leakage.

Finally, because True Negatives can be inflated by the abundance of "alert" samples, we used PR-AUC as the primary metric. PR-AUC heavily penalizes false alarms, reflecting the system's capacity to minimize "alarm fatigue" in actual deployments [33].

IV. Results

A. Performance of Single-Modality Baselines

We first evaluated single-modality model performance under the LOSO cross-drivers protocol. As shown in Table I, when using only RESP, all classifiers struggled to perform reliably across individuals, with PR-AUC ranging from 0.140 to 0.171, only marginally above the chance-level baseline (0.119), indicating that respiratory features alone were insufficient as standalone fatigue indicators.

The cardiovascular modality showed stronger predictive capability, with ECG-only PR-AUC ranging from 0.199 (Random Forest) to 0.210 (SVM). Although HRV features captured part of the variance in fatigue, absolute performance remained limited, highlighting the vulnerability of ECG-only systems to individual differences and physiological noise in a naturalistic driving environment.

B. Performance of Multimodal Fusion

Integrating ECG and RESP features improved model performance across all evaluated classifiers (Fig. 5), with the largest gain in tree-based ensemble models, i.e., the RF model. Specifically, for RF models, ECG+RESP achieved the highest PR-AUC of 0.249. Compared with the ECG-only baseline for the same model (0.199), fusion yielded an absolute improvement of 0.051, corresponding to a **25.5% relative increase**. XGBoost improved from 0.208 to 0.224, while KNN and SVM showed smaller but consistent gains. These results indicate that cardiovascular and respiratory signals carry complementary information for fatigue detection. Note that the values in Table I are descriptive LOSO estimates; a significance test was not performed because only eight subject-level folds were available.

C. Effects of Multimodal Fusion on False Alarm Reduction

For the best model (i.e., RF), we observed varying effects of multimodal fusion on PR-AUC and ROC-AUC. For the RF model, adding RESP increased PR-AUC from 0.199 to 0.249, whereas ROC-AUC remained nearly unchanged (0.611 to 0.616). Because 88.1% of samples were labeled as alert, ROC-AUC was less sensitive to false positives under this class imbalance. In contrast, PR-AUC focused on the sparse reduced-alertness class and was strongly affected by prediction precision. The higher PR-AUC, therefore, suggests that ECG+RESP fusion may have reduced some ECG-only false positives and made positive predictions more reliable. This is practically important because unnecessary warnings can reduce drivers' trust in the DMS.

V. Discussion

A. Physiological Synergy in Long-Haul Driving

The superiority of ECG+RESP fusion over single modalities offers insights into a key question in driver fatigue monitoring in naturalistic long-haul driving: how to decouple fatigue-related changes from subject-specific baseline shifts due to time-on-task effects. HRV features extracted from ECG are sensitive indicators of autonomic nervous system (ANS) transitions. However, over a 10-hour driving period, these HRV features may shift systematically, partially because prolonged sedentary posture can lower resting HR and alter parasympathetic tone, as reflected in RMSSD, independently of actual sleepiness [34]. Single-modality classifiers might therefore misinterpret natural physiological relaxation during long-haul drives as drowsiness, consistent with the performance ceiling and false-alarm tendency observed in ECG-only results [12].

However, the results also show that ECG carries rich fatigue-related information as a standalone signal: ECG-only PR-AUC consistently exceeded RESP-only PR-AUC. Thus, the gain from feature fusion should not be interpreted as respiration replacing ECG. Rather, it is possible that RESP improved precision by rejecting a subset of ECG-only false positives. While less sensitive to acute fluctuations in alertness, respiratory rate and amplitude are relatively stable over time,

TABLE I: Cross-Driver Performance Comparison by Modality under LOSO Evaluation

Classifier	PR-AUC			ROC-AUC			PR-AUC Gap
	ECG Only	RESP Only	**ECG+RESP**	ECG Only	RESP Only	**ECG+RESP**	(Fusion vs. ECG)
Random Forest	0.199	0.155	**0.249**	0.611	0.470	**0.616**	**+0.051**
XGBoost	0.208	0.140	**0.224**	0.598	0.470	**0.589**	+0.016
KNN ($k = 7$)	0.203	0.171	**0.212**	0.605	0.484	**0.580**	+0.008
SVM (RBF)	0.210	0.154	**0.217**	0.623	0.497	**0.602**	+0.007

Note: The chance-level baseline for PR-AUC is 0.119. Gap represents the absolute PR-AUC improvement over the ECG-only baseline.

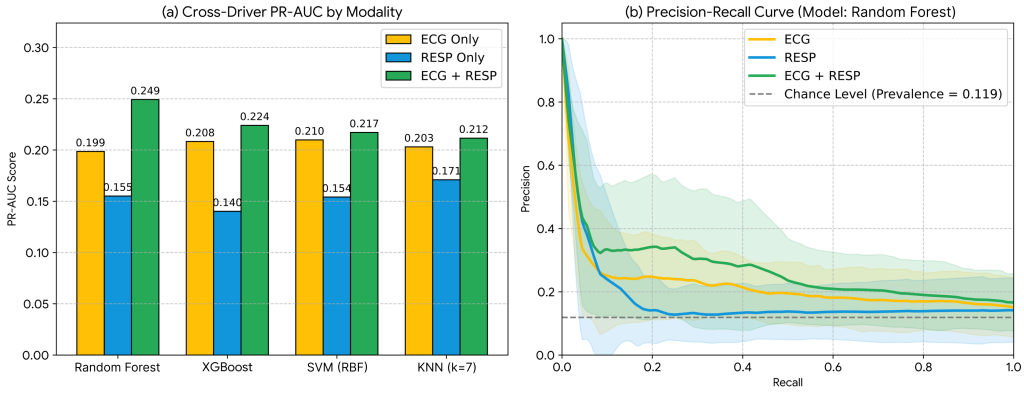


Fig. 5: Performance comparison of single-modality and multimodal models. (a) Cross-drivers PR-AUC across four classifiers. (b) Precision-Recall curves for the Random Forest model, highlighting the superior performance of ECG+RESP fusion.

and can help cross-validate cardiac anomalies. Fusion, therefore, leveraged physiological coupling, triggering fatigue alerts only when cardiac changes were corroborated by respiratory suppression, such as when shallower or slower breathing appeared [35].

B. Implications for Industrial Deployments

In commercial heavy-duty logistics, estimation accuracy can directly affect whether drivers trust and use a DMS. Fatigue-related events are sparse, accounting for only 11.9% of samples in our dataset. Thus, systems that emphasize ROC-AUC may appear reliable while still producing many false-positive warnings. Repeated false warnings can reduce users' trust in and compliance with warning systems, a phenomenon commonly referred to as the cry-wolf effect [36]. By improving PR-AUC from 0.199 to 0.249, ECG+RESP fusion increased the likelihood that an alert corresponds to a true reduced-alertness event. This precision gain is practically important for fleet deployment, not merely an algorithmic improvement.

C. Limitations and Future Work

While this study was based on a 160-hour naturalistic dataset, several limitations remain. First, although the per-driver duration was extensive (∼20 hours), the cohort included only eight professional male drivers. Future research should validate these multimodal synergies across larger and more diverse driver populations.

Second, the hourly KSS provided only sparse labels for driver state. Although backward mapping avoided temporal leakage, it assumed local stability within the preceding hour and might obscure fine-grained micro-sleep dynamics. Further, although we have reduced the KSS probe frequency to hourly, the subjective questionnaire may still bias the natural accumulation of fatigue. Future datasets should consider objective behavioral measures such as eyelid closure, gaze, head movement, and lane-departure events as ground truth of alertness.

Finally, the current study used classical classifiers and static features extracted from 5-minute windows. Future work should examine how personalized thresholds and ADAS active/inactive states can affect model performance. Models that consider temporal information, such as Recurrent Neural Networks or Transformers, should also be adopted to capture temporal physiological progression.

VI. Conclusion

Fatigue detection based on physiological signals often fails in real-world environments because of the lack of a naturalistic dataset and significant individual differences among drivers. In this study, using a 160-hour naturalistic long-haul truck dataset and a LOSO cross-validation protocol, we demonstrated the limitations of a single-modality model for detecting driver fatigue. Our results show that fusing ECG and RESP features yields better performance than models using a single modality. By fusing respiratory and cardiovascular signals, the multimodal RF classifier achieved a PR-AUC of 0.249, a 25.5% relative improvement over the best single-modality baseline, indicating that the multimodal physiological fusion

can help mitigate alarm fatigue and support more reliable fatigue detection in continuous driving.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the logistics industry partner for supporting this study, the engineers for assisting in-cabin data collection and the professional truck drivers for their participation. This research was funded by the Nansha District Key Area S&T Scheme, Science and Technology Bureau of Nansha District (No. 2023ZD006), and the Natural Science Foundation of Guangdong Province of China (No. 2024A1515010392).

REFERENCES

[1] M. K. Lemke, Y. Apostolopoulos, A. Hege, S. Sönmez, and L. Wideman, "Understanding the role of sleep quality and sleep duration in commercial driving safety," *Accident Analysis & Prevention*, vol. 97, pp. 79–86, 2016.

[2] M. M. Mitler, J. C. Miller, J. J. Lipsitz, J. K. Walsh, and C. D. Wylie, "The sleep of long-haul truck drivers," *New England Journal of Medicine*, vol. 337, no. 11, pp. 755–762, 1997.

[3] H. Häkkänen and H. Summala, "Fatal traffic accidents among trailer truck drivers and accident causes as viewed by other truck drivers," *Accident Analysis & Prevention*, vol. 33, no. 2, pp. 187–196, 2001.

[4] A. S. Aghaei, B. Donmez, C. C. Liu, D. He, G. Liu, K. N. Plataniotis, H.-Y. W. Chen, and Z. Sojoudi, "Smart driver monitoring: When signal processing meets human factors: In the driver's seat," *IEEE Signal Processing Magazine*, vol. 33, no. 6, pp. 35–48, 2016.

[5] C. C. Liu, S. G. Hosking, and M. G. Lenne, "Predicting driver drowsiness using vehicle measures: Recent insights and future challenges," *Journal of Safety Research*, vol. 40, no. 4, pp. 239–245, 2009.

[6] Z. Li, S. E. Li, R. Li, B. Cheng, and J. Shi, "Online detection of driver fatigue using steering wheel angles for real driving conditions," *Sensors*, vol. 17, no. 3, p. 495, 2017.

[7] T. Akerstedt and M. Gillberg, "Subjective and objective sleepiness in the active individual," *International Journal of Neuroscience*, vol. 52, no. 1-2, pp. 29–37, 1990.

[8] K. Kaida, M. Takahashi, T. Akerstedt, A. Nakata, Y. Otsuka, T. Haratani, and K. Fukasawa, "Validation of the karolinska sleepiness scale against performance and eeg variables," *Clinical Neurophysiology*, vol. 117, no. 7, pp. 1574–1581, 2006.

[9] M. L. Jackson, S. Raj, R. J. Croft, A. C. Hayley, L. A. Downey, G. A. Kennedy, and M. E. Howard, "Slow eyelid closure as a measure of driver drowsiness and its relationship to performance," *Traffic Injury Prevention*, vol. 17, no. 3, pp. 251–257, 2016.

[10] M. Ingre, T. Akerstedt, B. Peters, A. Anund, and G. Kecklund, "Subjective sleepiness, simulated driving performance and blink duration: examining individual differences," *Journal of Sleep Research*, vol. 15, no. 1, pp. 47–53, 2006.

[11] D. He, B. Donmez, C. C. Liu, and K. N. Plataniotis, "High cognitive load assessment in drivers through wireless electroencephalography and the validation of a modified n-back task," *IEEE Transactions on Human-Machine Systems*, vol. 49, no. 4, pp. 362–371, 2019.

[12] K. Lu, A. S. Dahlman, J. Karlsson, and S. Candefjord, "Detecting driver fatigue using heart rate variability: A systematic review," *Accident Analysis & Prevention*, vol. 178, p. 106830, 2022.

[13] J. Wang, H. Lu, A. Wang, X. Yang, Y. Chen, D. He, and K. Wu, "Physmle: Generalizable and priors-inclusive multi-task remote physiological measurement," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[14] H. Mårtensson, O. Keelan, and C. Ahlström, "Driver sleepiness classification based on physiological data and driving performance from real road driving," *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 2, pp. 421–430, 2018.

[15] Y. Jiao, C. Zhang, X. Chen, L. Fu, C. Jiang, and C. Wen, "Driver fatigue detection using measures of heart rate variability and electrodermal activity," *IEEE Transactions on Intelligent Transportation Systems*, vol. 25, no. 6, pp. 5510–5524, 2023.

[16] H. S. Stern, D. Blower, M. L. Cohen, C. A. Czeisler, D. F. Dinges, J. B. Greenhouse, F. Guo, R. J. Hanowski, N. P. Hartenbaum, G. P. Krueger *et al.*, "Data and methods for studying commercial motor vehicle driver fatigue, highway safety and long-term driver health," *Accident Analysis & Prevention*, vol. 126, pp. 37–42, 2019.

[17] A. Anund, C. Fors, D. Hallvig, T. Akerstedt, and G. Kecklund, "Observer rated sleepiness and real road driving: An explorative study," *PLoS ONE*, vol. 8, no. 5, p. e64782, 2013.

[18] R. J. Hanowski, W. W. Wierwille, and T. A. Dingus, "An on-road study to investigate fatigue in local/short haul trucking," *Accident Analysis & Prevention*, vol. 35, no. 2, pp. 153–160, 2003.

[19] R. Friswell and A. Williamson, "Comparison of the fatigue experiences of short haul light and long distance heavy vehicle drivers," *Safety Science*, vol. 57, pp. 203–213, 2013.

[20] E. Perkins, C. Sitaula, M. Burke, and F. Marzbanrad, "Challenges of driver drowsiness prediction: The remaining steps to implementation," *IEEE Transactions on Intelligent Vehicles*, vol. 8, no. 2, pp. 1319–1338, 2022.

[21] C. Ahlström and A. Anund, "Development of sleepiness in professional truck drivers: Real-road testing for driver drowsiness and attention warning (DDAW) system evaluation," *Journal of Sleep Research*, vol. 34, no. 2, p. e14259, 2025.

[22] B. Bakker, B. Zabłocki, A. Baker, V. Riethmeister, B. Marx, G. Iyer, A. Anund, and C. Ahlström, "A multi-stage, multi-feature machine learning approach to detect driver sleepiness in naturalistic road driving conditions," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 5, pp. 4791–4800, 2021.

[23] C. Ahlström, W. van Leeuwen, S. Krupenia, H. Jansson, S. Finér, A. Anund, and G. Kecklund, "Real-time adaptation of driving time and rest periods in automated long-haul trucking: Development of a system based on biomathematical modelling, fatigue and relaxation monitoring," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 5, pp. 4758–4766, 2021.

[24] G. Sikander and S. Anwar, "Driver fatigue detection systems: A review," *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 6, pp. 2339–2352, 2018.

[25] S. O.-R. A. V. S. Committee *et al.*, "Taxonomy and definitions for terms related to on-road motor vehicle automated driving systems," *SAE Standard J*, vol. 3016, p. 1, 2014.

[26] C. Ahlström, R. Zemblys, H. Jansson, C. Forsberg, J. Karlsson, and A. Anund, "Effects of partially automated driving on the development of driver sleepiness," *Accident Analysis & Prevention*, vol. 153, p. 106058, 2021.

[27] A. Å. Miley, G. Kecklund, and T. Åkerstedt, "Comparing two versions of the Karolinska Sleepiness Scale (KSS)," *Sleep and Biological Rhythms*, vol. 14, no. 3, pp. 257–260, 2016.

[28] D. Makowski, T. Pham, Z. J. Lau, J. C. Brammer, F. Lespinasse, H. Pham, C. Schölzel, and S. H. A. Chen, "NeuroKit2: A Python toolbox for neurophysiological signal processing," *Behavior Research Methods*, vol. 53, no. 4, pp. 1689–1696, 2021.

[29] J. Cook and V. Ramadas, "When to consult precision-recall curves," *The Stata Journal*, vol. 20, no. 1, pp. 131–148, 2020.

[30] S. H. Khan, M. Hayat, M. Bennamoun, F. A. Sohel, and R. Togneri, "Cost-sensitive learning of deep feature representations from imbalanced data," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 8, pp. 3573–3587, 2017.

[31] I. Araf, A. Idri, and I. Chairi, "Cost-sensitive learning for imbalanced medical data: a review," *Artificial Intelligence Review*, vol. 57, no. 4, p. 80, 2024.

[32] H. Ney, U. Essen, and R. Kneser, "On the estimation of 'small' probabilities by leaving-one-out," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 17, no. 12, pp. 1202–1212, 2002.

[33] T. Saito and M. Rehmsmeier, "The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets," *PLoS ONE*, vol. 10, no. 3, p. e0118432, 2015.

[34] F. Shaffer and J. P. Ginsberg, "An overview of heart rate variability metrics and norms," *Frontiers in Public Health*, vol. 5, p. 290215, 2017.

[35] J. Solaz, J. Laparra-Hernández, D. Bande, N. Rodríguez, S. Veleff, J. Gerpe, and E. Medina, "Drowsiness detection based on the analysis of breathing rate obtained from real-time image recognition," *Transportation Research Procedia*, vol. 14, pp. 3867–3876, 2016.

[36] F. Naujoks, A. Kiesel, and A. Neukum, "Cooperative warning systems: The impact of false and unnecessary alarms on drivers' compliance," *Accident Analysis & Prevention*, vol. 97, pp. 162–175, 2016.