A Multi-modal Drowsy Driver Detection with Random Forest using Video-based and Physiological Features

Steven Sison[†*], Miguel Aldo Valbuena[*], Ron Louis Nierva[*], Kathleen Issandra Tuso[*], Jon Dewitt Dalisay, PhD[†]
[†]*Artificial Intelligence Program*, [*]*Electrical and Electronics Engineering Institute*,
University of the Philippines Diliman, Quezon City, Philippines
{sssison1, mavalbuena, rcnierva, kttuso, jedalisay}@up.edu.ph

Abstract—**Driver drowsiness is a common occurrence in road accidents, affecting road safety. This study proposes a hybrid multi-modal data fusion approach that combines facial video features with electroencephalography (EEG) and electrocardiogram (ECG) signals to improve the accuracy and robustness of drowsy driver detection. Using the DROZY dataset, our system extracts spatio-temporal features from both video and physiological data, then merges these features to improve accuracy. The resulting dataset was used to train a model utilizing the random forest algorithm (RF) as the classifier and principal component analysis (PCA) with 95% explained variance for dimensionality reduction. The results show that a higher performance was achieved by combining video-based features and physiological information from EEG and ECG; peaking at around 93%-94% in all performance metrics with the multi-modal approach. Moreover, the effect of varying window sizes (5s, 10s, 30s, 60s) was also investigated, wherein the 30s window generally showed the optimal performance results, while only being edged by the 10s window under the multi-modal scheme. To combat the loss of temporal resolution in increasing window sizes, a sliding window technique may be applied in future iterations of multi-modal DDD research, finding a balance between resolution and computational complexity trade-offs.**

Index Terms—**drowsiness detection, machine learning, physiological signals, computer vision, multi-modal data fusion**

I. Introduction

Road accidents share a large portion of deaths globally. These incidents lead to fatalities, injuries, and infrastructural damages. In the Philippines, road crash accidents increased by 35% from 2023 to 2024, majority due to reckless driving [1]. This increase in casualties necessitates the development of risk prevention systems for driving, aside from policies.

One of the hazards connected to reckless driving is driver drowsiness. Driver alertness is an essential element for road safety, as fatigued drivers may suffer from impaired senses, slow reaction time, and poor muscle coordination which becomes a hazard [2]. To combat its occurrence, research on driver drowsiness detection (DDD) systems have been continuously conducted, providing timely warnings to drivers. Some applications include intelligent driving systems that take control of vehicles when drivers become drowsy; some integrate comfort systems according to the DDD result [3]–[5].

Traditional approaches in DDD include the use of vision-based and physiological signals data. The paper from [6] presented a driver drowsiness estimation system using electroencephalogram (EEG) signals through an encoder-decoder network to determine the percentage of eyelid closure over pupil over time (PERCLOS). The model in their work used the spectral features of decomposed EEG sub-bands across eight channels. While the use of physiological signals for DDD prove to be a viable approach, physiological workloads are shown to have higher and more irregular values during real driving conditions compared to simulated conditions [7]. As such, it is still advisable to use other hybrid measures for DDD to ensure reliability and robustness.

By the advances in computer vision, camera-based approaches have been gaining popularity for DDD. Compared to the earlier approach, it is non-invasive and requires only a camera sensor.

Most vision-based approaches use a tracking algorithm to track facial landmarks such as the eyes and mouth to measure eye blinking, eye aspect ratio, facial expressions, yawning, head pose, and gaze. Since these features are key indicators of drowsiness, several studies utilizing these features show great predictive performance [8]–[10]. However, similar to the physiological signal-based approach, real driving conditions are harsher than simulated driving conditions. Low lighting conditions prove to be a challenge when using vision-based approaches due to the difficulty in detecting facial landmarks. Additionally, obstructions such as glasses can also impede the detection of facial features.

This work proposes a multi-modal driver drowsiness detection framework that combines facial video features with EEG and ECG signals. This study is further motivated by biosignal acquisition systems such as BrainFlow and OpenBCI, which have enabled gathering of physiological data using wearable devices [11], [12]. This study aims to further explore the multi-modal data fusion approach, incorporating CV, ECG, and EEG signals in a machine learning framework for DDD. Using the DROZY dataset, our system merges the spatio-temporal features from both modalities to improve accuracy. We apply a Random Forest to determine drowsiness levels and explore how different time windows (5s, 10s, 30s, 60s) affect detection performance. Inference efficiency and key features were evaluated to ensure practical viability and interpretability.

II. Methodology

This section outlines the methodology adopted for hybrid multi-modal DDD. The overall workflow is illustrated in Figure 1, which involves data preprocessing using DROZY dataset, feature engineering through extraction and fusion,

979-8-3315-6946-4/25/$31.00 ©2025 IEEE

model training using a Random Forest classifier, and performance evaluation.

A. Dataset Description

The ULg multi-modality Drowsiness Database or DROZY was used in this study [13]. It consists of 14 different subjects each subjected to three (3) successive tests where time-synchronized video sensor data and polysomnography signals (EEG, EOG, ECG, and EMG) were obtained. In this study, we utilized the video sensor data, EEG signal, and ECG signal to detect driver drowsiness. The videos have a dimension of 512x512 in 8-bit scale with 15/30 FPS while the PSG signals contain 5-channel EEG and ECG sampled at 512 Hz. The dataset used the Karolinska Sleepiness Scale (KSS) scores for indicating drowsiness, but was dichotomized to a binary value setting 1 as drowsy and 0 as non-drowsy following the scale of [14] where KSS scores of less than or equal to six (6) were set as non-drowsy while the remaining scores were set to drowsy.

B. Feature Engineering

The drowsy driver detection was treated as a spatio-temporal problem, wherein spatial features such as facial landmarks in a video were obtained and temporal changes in these facial landmarks were recorded and observed. To facilitate the extraction of spatiotemporal features, each 10-min test, having video and corresponding physiological data, was segmented into nonoverlapping smaller temporal window chunks (5s, 10s, 30s, and 60s). This chunking strategy allows for better localization of features that allows for the detection of behavioral patterns in both video (blinking, head pose, gaze) and physiological (heart rate variability) data that may be undetectable in longer time frames. Each chunk is treated as an independent data point and will have the same class label as the 10-min test it was extracted from.

1) Video-Based Features: Video-based features are obtained by using a face detector to obtain facial landmarks. Most video-based features rely on facial features, such as the eyes and mouth, to identify drowsiness. This study focuses only on the spatiotemporal changes in the subjects' eyes [8], [15], head orientation [16], and gaze [17]. The detailed explanation regarding the video-based features were summarized in Table I.

TABLE I
VIDEO-BASED FEATURES

Feature	Description
Eye Aspect Ratio	Distances among eye landmarks
Blinks	Number of blinks over a window size
Blink Length	Width of peaks at half-prominence level
Closing Blink Velocity	Closing speed during blinking.
Opening Blink Velocity	Opening speed during blinking.
Fixed Gaze Short	Determines still gaze
Fixed Gaze Score	Determines gaze fixation
Pitch	Up-and-Down orientation of the head.
Yaw	Left-to-right orientation of the head.
Roll	Side-to-side tilt of the face.

2) Physiological Features: The .edf files from the DROZY dataset contain time-series ECG and EEG signals that were recorded in drowsiness studies consisting of 14 test subjects, each with 5 electrode channels. From there, features were extracted in n-second epochs, where n = 5, 10, 30, and 60 seconds. Four main families of features are derived: Power Spectral Density (PSD), Time-Domain, Nonlinear, and ECG-based features. All features identified in the .edf files are summarized in Table II.

TABLE II
EEG AND ECG FEATURES

Feature	Type	Description
bpd	PSD	Band power in Delta band (0.5–4Hz)
bpt	PSD	Band power in Theta band (4–8Hz)
bpa	PSD	Band power in Alpha band (8–13Hz)
bpb	PSD	Band power in Beta band (13–30Hz)
bpg	PSD	Gamma Band Power (30-100Hz)
rba	PSD	Ratio of Alpha-to-Beta power
re	PSD	Relative energy in frequency band
mean, md, me	Time	Mean, Median, and Mean energy
sd, var	Time	Standard deviation and variance
skew, kurt	Time	Skewness and kurtosis
1d, n1d	Time	First derivative, normalized form
2d, n2d	Time	Second derivative, normalized form
am	Time	Amplitude modulation
am	Nonlinear	Amplitude modulation
hm	Nonlinear	Hjorth Mobility
hc	Nonlinear	Hjorth Complexity
te, mte	Nonlinear	Teager Energy, Mean Teager Energy
lrssv	Nonlinear	Log Root Sum Square Value
mcl	Nonlinear	Mean Curve Length
Heart rate	ECG	Beats per minute
pNN50	ECG	Percent of RR intervals $>$ 50 ms
RMSSD	ECG	RMS of Successive Differences
SDNN	ECG	Standard deviation of NN intervals
mean_RR	ECG	Mean of RR intervals
VLF_power	ECG	Power in Very Low Frequency Band

C. Model Training and Evaluation

The resulting dataset was used to train a model utilizing the random forest algorithm (RF) as the classifier and principal component analysis (PCA) with 95% explained variance for dimensionality reduction. Baseline models utilizing only video-based features and physiological features were also created as a benchmark to determine the effectiveness of utilizing multimodal data for drowsy driver detection. All models were subjected to hyperparameter tuning using Bayesian optimization using the hyperparmeters outlined in Table III. The model was trained using K-fold cross validation with $K = 5$ to ensure better generalization against unseen data. The resulting models were evaluated in terms of their accuracy, precision, recall, F1-score, inference time, and memory consumption.

TABLE III
RANGE OF VALUES FOR HYPERPARAMETERS (RANDOM FOREST)

Hyperparameter	Values
n_estimators	Integer(10,100)
criterion	[gini, entropy]
max_depth	Integer(10,50)
min_samples_split	Integer(2,20)
min_samples_leaf	Integer(1,20)
max_features	[sqrt, log2]
bootstrap	[True, False]

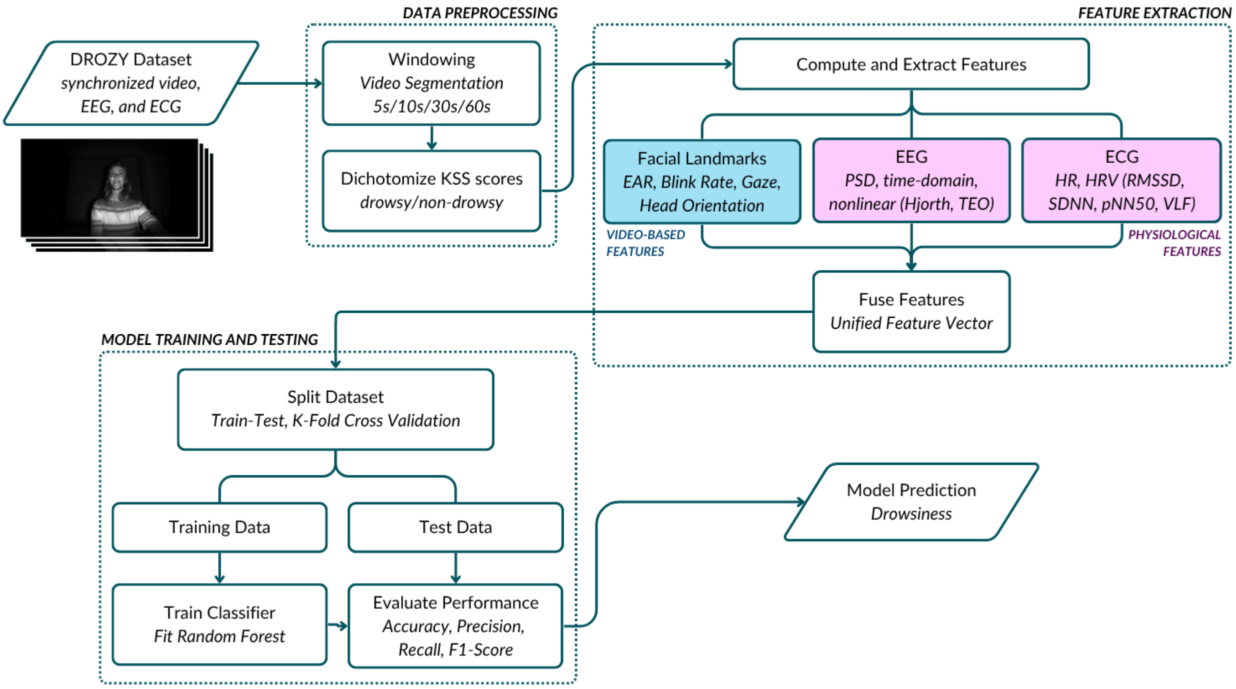

Fig. 1. Flowchart of the proposed hybrid multi-modal driver drowsiness detection system

III. Results and Discussion

A. Exploratory Data Analysis

1) Video-Based: Figure 2 shows the 30-second EAR signal and head pose estimate for subject 8 test 1 and test 3 which is labeled as non-drowsy and drowsy, respectively. The figure shows that non-drowsy drivers tend to have less number of blinks, longer blink length, and higher PERCLOS. In terms of head pose, non-drowsy drivers tend to maintain their head pose over longer periods while drowsy drivers have more sudden changes or spikes.

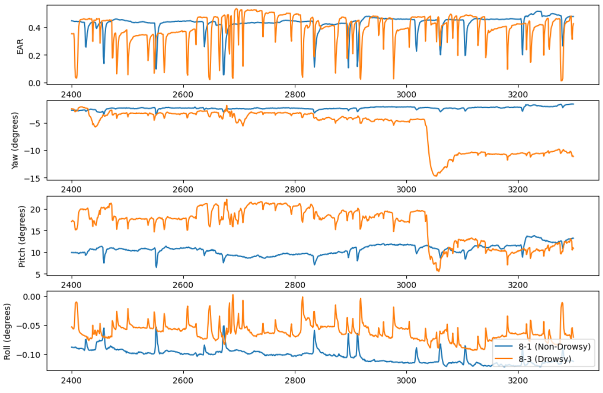

Fig. 2. Eye Aspect Ratio and Head Pose Estimate over Time

Figure 3 compares the FGScore over time using the same test subjects. Non-drowsy drivers have lower eye position error indicating stable and accurate gaze tracking, whereas drowsy drivers have shown several spikes in eye position error likely corresponding to moments of eye closure or erratic movement. Additionally, non-drowsy drivers tend to have an increasing FGScore, indicating consistent eye movement and alert attention, while drowsy drivers are more inconsistent suggesting prolonged fixation.

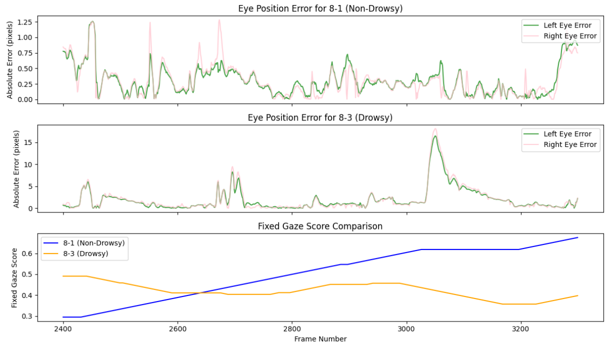

Fig. 3. Fixed Gaze Score over Time

2) Physiological: Among the physiological data, the ECG features contained heart rate information. The values for mean R-R and heart rate features had a linear correlation, which was expected as both were derived from the average distance between R-peaks. The RMSSD and SDNN values also had a linear relationship as both pertain to the variation of successive R-peak timestamps.

Upon inspection of the ECG features, data that indicate a higher heart rate and HRV tends to be classified as non-drowsy. The mean R-R and heart rate features from Fig. 6 indicate a higher HRV as shown by the large sharp transitions and frequent value fluctuations.

Since there were 100+ EEG features, they cannot be fitted into a single correlogram heatmap. However, features with strong correlations (positively or negatively) with each other were noted. EEG features that had a high correlation were the spectral information, particularly the sub-band powers (alpha,

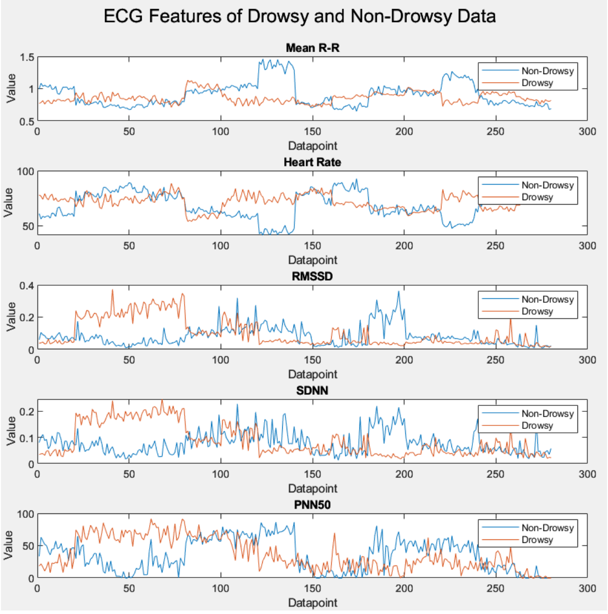

Fig. 4. ECG feature values for drowsy and non-drowsy labeled data points

beta, gamma, theta, and delta). This linear relationship was also observed within the PSD features and across all EEG channels. This may also imply that one channel is enough for drowsiness prediction.

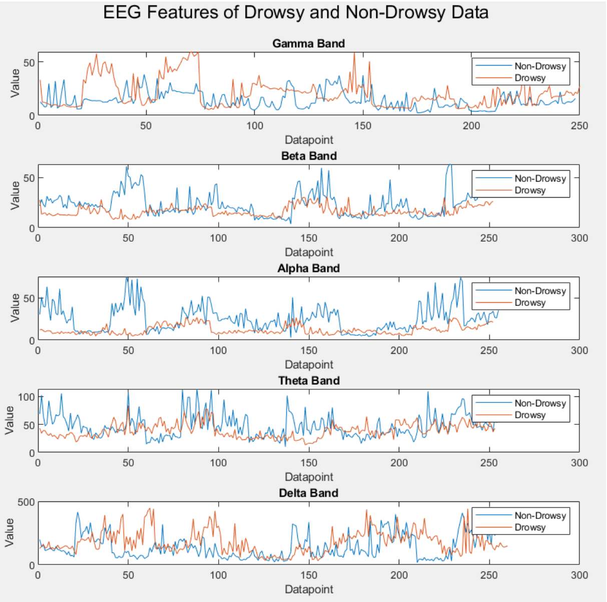

Fig. 5. EEG sub-bands power features for drowsy and non-drowsy labeled data points

As there was a high correlation of frequency band power across EEG channels, the Fz channel was used for exploratory data analysis. The power of the beta, alpha sub-bands for drowsy labeled data points are flatter compared to their non-drowsy counterparts. This corresponded to the description of the EEG frequency bands as mentioned earlier: alpha and beta are generally associated with alertness [18]. Gamma should also diminish with drowsiness, reflecting the loss of high-level cognitive activity, but the opposite is seen here.

The delta band also shows the opposite behavior: more fluctuations in drowsy than in the non-drowsy data points. Some non-drowsy participants have bursts in delta, but generally, delta activity is stronger in drowsy states. Delta waves dramatically increase in sleep and fluctuations may indicate transition to sleep [18].

B. *Effect of Different Window Size*

Table IV shows the performance comparison across different models trained on purely video features, purely physiological features, and multi-modal features with varying window size. Both baseline models using only purely video-based features and purely physiological features show an interesting trend, wherein increasing the window size improves the performance of the model only up to a certain point (up to 30s window size) to which the performance drops significantly when increased further. A similar trend was also observed when both features were combined. However, more importantly, the findings suggest that multi-modal models generally outperform models utilizing only either video-based or physiological feature sets.

In general, longer temporal windows can provide more information and improve the predictive accuracy of a model, but only up to a certain window size. When using video-based features, those obtained using short temporal windows often perform poorly due to poor contextualization. ECG and EEG signals, however, can have a small window size, as these essentially repeat over a short period of time only. However, as shown in the results, the 5s window size had worse performance than the rest. This may be due to the susceptibility of the feature extraction method to noise-corrupted data. ECG data from test subject 2-1 was heavily corrupted by noise such that the QRS complexes were not retrieved by the Pan-Tompkins algorithm. This eventually led to more data points in feature extraction when using smaller window sizes.

A previous study conducted by [19] observed similar findings in which the area under the ROC curve for drowsiness detection increases and saturates at a window size of 30s. This suggests that as long as the temporal window is long enough, the model will be able to obtain sufficient context and classify correctly. However, it is still also important to note that extremely large window sizes can still be detrimental to the model. If the temporal window is too long, overgeneralization might occur leading to some drowsiness episodes to be overlooked due to the averaging out some short-term indicators of drowsiness [20], [21]. Therefore, a moderate window size should be used in order to allow the model to contextualize enough while still preventing the possibility of an overgeneralization. This also ensures that the model is more robust to any perturbations or quick-state transitions (a drowsy-to-non-drowsy state in less than 60s) allowing the model to still detect these fast-paced changes in the drivers state.

Furthermore, looking at drowsiness detection in a real-time perspective also shows how vital it is to use the correct window size. Using a longer window size would indicate a longer time delay before drowsiness is detected, consequently increasing the time it takes to alert the driver. In contrast, a smaller window size allows for faster real-time detection of

TABLE IV
PERFORMANCE COMPARISON OF SPATIO-TEMPORAL FEATURES ACROSS DIFFERENT MODALITIES AND WINDOW SIZES.

Metric	**Video-Based**				**Physiological**				**Multi-modal**			
	5s	10s	30s	60s	5s	10s	30s	60s	5s	10s	30s	60s
Accuracy	0.8511	0.8657	0.8656	0.8148	0.8688	0.8889	0.8889	0.7870	0.9097	0.9398	0.9352	0.8981
Precision	0.8505	0.8660	0.8609	0.8228	0.8682	0.8887	0.8892	0.7855	0.9098	0.9400	0.9305	0.8902
Recall	0.8511	0.8657	0.8565	0.8148	0.8688	0.8889	0.8889	0.7870	0.9097	0.9398	0.9352	0.8981
F1-Score	0.8495	0.8640	0.8527	0.8071	0.8681	0.8881	0.8877	0.7827	0.9091	0.9395	0.9345	0.8968

drowsiness. Additionally, it might be better for future research to explore an overlapping window approach rather than a non-overlapping window approach to allow the system to handle incremental updates for lower latency in detection.

C. Inference Time and Memory Consumption

Figure 6 shows how the inference time and training time vary when using different types of features across different window sizes. As the window size increases, the training time decreases. This trend is expected since increasing the window size reduces the amount of chunks extracted from each test subjects. On the other hand, there is no visible trend as to how inference time varies within different window sizes. However, it can be observed that within the same window size, the difference among the prediction time of models using purely video-based features, purely physiological features, and multi-modal features is negligible.

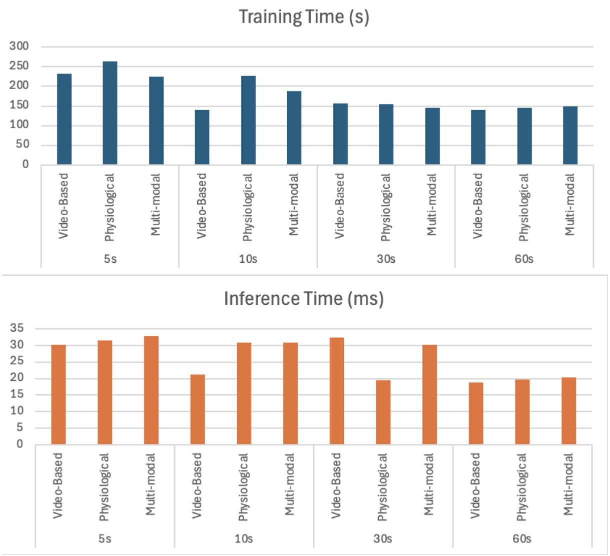

Fig. 6. Training Time and Inference Time

Additionally, Figure 7 shows the model size and memory consumption during inference for each model trained on different feature sets across different window sizes. The model size decreases as the window size is increased, particularly because of the same reason why the training time decreases, because of the decrease in number of chunks. On the other hand, the memory consumption during inference remain consistent within the same window sizes. This indicates that majority of the memory consumption during inference is primarily attributed to the features being stored rather than the actual usage of the model. Furthermore, the results show that at similar window sizes, there is a minimal difference between the memory consumption of the baseline models compared to the multi-modal model.

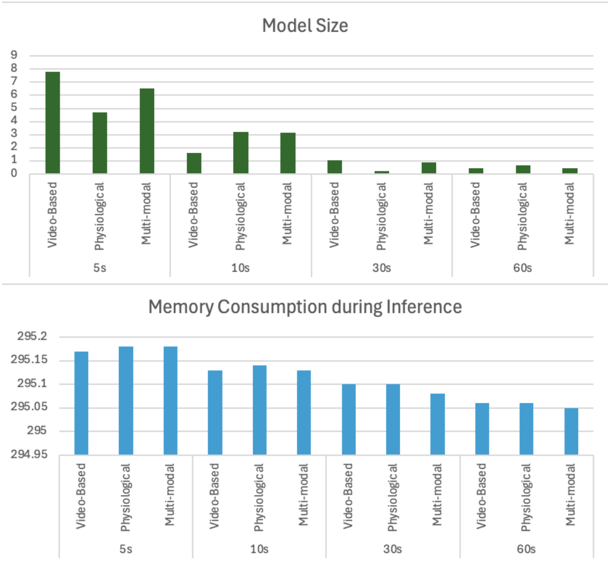

Fig. 7. Memory Consumption

D. Feature Importance

To ensure the explainability of the results, SHapley's additive explanations was also used to determine the top features contributing to the prediction of the model. For simplification, we look only at the best performing window-size (30s). Figure 8 shows the top 20 SHAP values for the resulting multi-modal model for detecting driver drowsiness. Results revealed a mix of both video-based and physiological-based features indicating the importance of using both modality in terms of drowsy driver detection.

IV. CONCLUSION

This study presented an evaluation of video-based, physiological, and multi-modal applications in drowsy driver detection. A higher performance was achieved by combining video-based features and physiological information from ECG and EEG; peaking at around 93%-94% in all performance metrics with the multi-modal approach. Moreover, the effect of varying window sizes was also investigated, wherein the 30s window size generally showed the optimal performance results, while only being edged by the 10s window under the multi-modal scheme.

The performance of the physiological modality model slightly increased as the window size went from 5 to 30s,

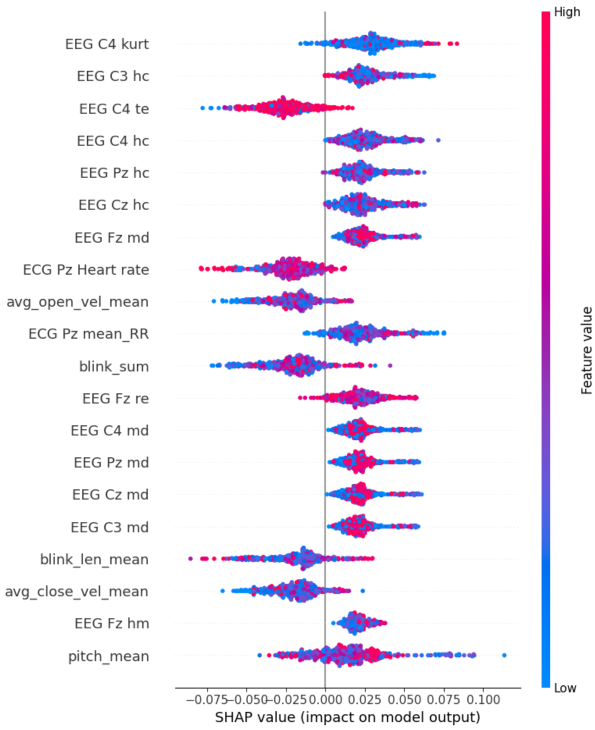

Fig. 8. Shapley Additive Values for Multi-modal Model trained on 30s Chunks

then decreased at 60s. Increasing the window size allowed for more temporal context to be included in the model, while larger window sizes may have also led to over-averaging the values from feature extraction, smoothing out the finer information within the signals. To combat the loss of temporal resolution in increasing window sizes, a sliding window technique may be applied in future iterations of multi-modal DDD research; although this can also lead to increase in computational complexity.

Aside from performance, increasing the window size also led to smaller model sizes and lower inference memory consumption. Though this may be beneficial for edge computing, the larger window size can imply having to wait longer for enough input data before processing. This implies that a larger window size can be detrimental in real-time applications such as DDD, wherein immediate feedback is crucial.

While the multi-modal approach produced the best results, the performance of using either solely a video-based or a solely physiological modality is comparable to each other. This may indicate that a similar accuracy performance can be achieved by using either approach. Additionally, this can potentially be a basis for deploying a data fusion-based robust DDD system where one modality can be used in the event that the other modality fails (i.e. using ECG and EEG if image-based data is unavailable or anomalous, vice versa), in addition to just combining all data sources.

References

[1] E. Tupas, "Road crash fatalities up 35% in 2024 – hpg," Feb. 2025. Accessed: 2025-02-02.

[2] S. Saleem, "Risk assessment of road traffic accidents related to sleepiness during driving: a systematic review," *East Mediterr Health J.*, vol. 28, no. 9, pp. 695–700, 2022. Received: 09/10/21; accepted: 11/05/22.

[3] 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, pp. 1319–1338, Feb. 2023.

[4] M. Ahmed, S. Masood, M. Ahmad, and A. A. A. El-Latif, "Intelligent driver drowsiness detection for traffic safety based on multi cnn deep model and facial subsampling," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, pp. 19743–19752, Oct. 2022.

[5] T. Horberry, "Human-centered design for an in-vehicle truck driver fatigue and distraction warning system," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, pp. 5350–5359, June 2022.

[6] S. Arefnezhad *et al.*, "Driver drowsiness estimation using eeg signals with a dynamical encoder–decoder modeling framework," *Scientific Reports*, vol. 12, p. 2650, Feb. 2022.

[7] A. Sahayadhas, K. Sundaraj, and M. Murugappan, "Detecting driver drowsiness based on sensors: a review," *Sensors (Basel)*, vol. 12, pp. 16937–16953, Dec. 2012. Erratum in: Sensors (Basel). 2021 Jan 11;21(2):E451. doi: 10.3390/s21020451.

[8] W. Deng and R. Wu, "Real-time driver-drowsiness detection system using facial features," *IEEE Access*, vol. 7, pp. 118727–118738, 2019.

[9] M. Ahmed, S. Masood, M. Ahmad, and A. Abd El-Latif, "Intelligent driver drowsiness detection for traffic safety based on multi cnn deep model and facial subsampling," *IEEE Transactions on Intelligent Transportation Systems*, vol. PP, pp. 1–10, 12 2021.

[10] M. Vijay, N. Vinayak, M. Nunna, and N. Subramanyam, "Real-time driver drowsiness detection using facial action units," pp. 10113–10119, 01 2021.

[11] BrainFlow, "Installation instructions." https://brainflow.readthedocs.io/en/stable. Accessed: 2025-05-30.

[12] "Welcome to the openbci community," Feb. 2022.

[13] Q. Massoz, T. Langohr, C. François, and J. G. Verly, "The ulg multimodality drowsiness database (called drozy) and examples of use," in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1–7, 2016.

[14] S. Puttonen, K. Karhula, A. Ropponen, T. Hakola, M. Sallinen, and M. Härmä, "Sleep, sleepiness and need for recovery of industrial employees after a change from an 8- to a 12-hour shift system," *Industrial Health*, vol. 60, 10 2021.

[15] S. Zainal, I. Khan, and H. Abdullah, "Efficient drowsiness detection by facial features monitoring," *Research Journal of Applied Sciences, Engineering and Technology*, vol. 7, pp. 2376–2380, 03 2014.

[16] I.-H. Choi and Y.-G. Kim, "Head pose and gaze direction tracking for detecting a drowsy driver," in *2014 International Conference on Big Data and Smart Computing (BIGCOMP)*, pp. 241–244, 2014.

[17] L. M. Bergasa, J. Nuevo, M. A. Sotelo, R. Barea, and M. E. Lopez, "Real-time system for monitoring driver vigilance," *IEEE Transactions on Intelligent Transportation Systems*, vol. 7, pp. 63–77, Mar. 2006.

[18] Author(s), "Eeg signal processing techniques and applications," *Sensors*, vol. 23, no. 22, p. 9056, 2023.

[19] E. M. Vural, M. Cetin, G. Littlewort, M. Bartlett, and J. Movellan, "Automated drowsiness detection for improved driving safety," in *International Conference on Automotive Technologies (ICAT)*, pp. 1–6, 2008.

[20] R. Ghoddoosian, M. Galib, and V. Athitsos, "A realistic dataset and baseline temporal model for early drowsiness detection," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 178–187, 2019.

[21] Y. Albadawi, A. AlRedhaei, and M. Takruri, "Real-time machine learning-based driver drowsiness detection using visual features," *Journal of Imaging*, vol. 9, no. 5, 2023.