On-Device Multimodal Drowsy Vehicle Driver Detection and Alerting System

Jong-seok Yoon*, Youhyeon Choi*, Yiwen Shen*
* *Department of Software and Computer Engineering*,
Ajou University,
Suwon, Republic of Korea
whdtjr22@ajou.ac.kr, nanana126@ajou.ac.kr, chrisshen@ajou.ac.kr

***Abstract*—Traffic accidents caused by health-related issues are difficult to respond to promptly, particularly during long-duration driving, where risks such as drowsiness and cardiovascular abnormalities increase. In this study, we introduce a multimodal state recognition system capable of detecting drowsiness and health anomalies, using visual and physiological signals collected from a camera and a PPG sensor. We constructed the time-series data using frame-level features such as PERCLOS, MAR, and PPG signals, and compared the performances across various deep learning architectures, including an attention-based Transformer model. Our experiment results shows that the proposed approach achieves higher detection accuracy compared to other traditional models. Additionally, the system supports active intervention through voice alerts and automatic emergency reporting with GPS location information in hazardous situations. We implemented the proposed system in a low-cost on-device environment, which demonstrates its feasibility for real-world applications.**

***Index Terms*—Drowsiness, multimodality, on-device, real-time**

I. Introduction

Traffic accidents caused by drowsiness and health-related issues during driving have become a critical social problem [1], with over 6,300 cases reported annually in South Korea alone. These incidents not only endanger the driver but also pose serious risks to surrounding individuals. The risk is particularly high for elderly drivers or those engaged in prolonged driving, where drowsiness may coincide with cardiovascular abnormalities. Fig. 1 shows the potential dangers associated with drowsy driving. However, current vehicle systems face limitations in rapid emergency response, often due to delays in sending rescue signals after an incident [2]–[4].

Conventional drowsiness detection systems primarily rely on camera-based visual cues or basic physiological sensors, and their functionality is often limited to passive approaches such as triggering simple audible alarms after detecting signs of fatigue. Furthermore, given the driving context where hands-free interaction is essential, current systems lack active user interaction or multimodal integration.

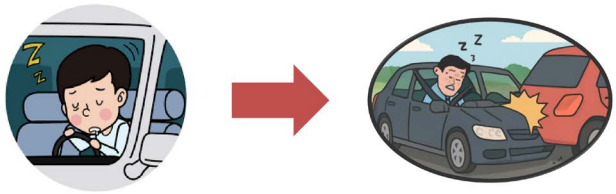


Fig. 1. The danger of drowsy driving.

Several prior works have explored the use of multimodal approaches in drowsiness detection. Yu et al. employed facial features, head posture, and PPG signals with LSTM-based models, forming 10-dimensional fused vectors using BiLSTM [5]. However, scalability and real-time deployment remained limited. Du et al. used a late fusion strategy combining PPG and video-based models, but failed to capture inter-modal correlations [6]. Cao et al. proposed a multimodal approach using ECG (i.e. Electrocardiogram), EEG (i.e. Electroencephalogram), and EMG (i.e. Electromyography), but required expensive sensors and complex setups, making them unsuitable for embedded deployment [7].

Unlike LSTM-based methods that suffer from long-term dependency loss and limited parallelization, the Transformer architecture built on attention mechanisms can simultaneously capture global temporal relationships and is optimized for parallel processing. While applications of Transformers in drowsiness detection remain rare, this study explores its feasibility and advantages in an embedded context.

To address these limitations, we propose a multimodal drowsiness and health anomaly detection system that operates in a low-cost embedded environment using Jetson Nano and Raspberry Pi. The system integrates frame-based visual features such as PERCLOS (i.e. PERcentage of eye CLOSure) and MAR (i.e. Mouth Aspect Ratio) with HRV (i.e. Heart Rate Variability)-based physiological signals acquired via a PPG (i.e. Photoplethysmography) sensor. These signals are merged at the feature level using an early fusion strategy and converted into time-series data. We then apply and compare multiple deep learning models, including CNN, LSTM, and Transformer, to identify the optimal architecture for accurate

This research was supported in part by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2022R1I1A1A01053915), in part by the MSIT (Ministry of Science and ICT), Korea, under the National Program for Excellence in SW (2022-0-01077) supervised by the IITP (Institute of Information & communications Technology Planning & Evaluation) in 2025, and in part by Ajou University research fund. *(Jong-seok Yoon and Youhyeon Choi are co-first authors.) (Corresponding author: Yiwen Shen.)*

and efficient state recognition. The proposed system using a Transformer-based early fusion model to achieve up to 79.35% accuracy of drowsiness detection. To provide an in-time notification, we also integrate an alert generation, a conversational interaction, and a GPS-based reporting function in a seamless flow for detecting drowsiness of a vehicle driver. In hazardous cases, the alerting module issues audio alerts and automatically transmits the driver's location and state to a server, significantly reducing emergency response time.

The key technical contributions are summarized as follows:

- In the proposed system, we integrate visual features (PERCLOS, MAR, EAR) and physiological/biological signals (PPG) via time-series feature-level early fusion.
- We use the Transformer with 1-D convolution to demonstrate that the proposed system outperforms other state-of-the-art models in accuracy.
- We also implement audio alerts and GPS-based automatic emergency reporting, reducing response time in critical scenarios.
- For an edge deployment scenario, we develop a on-device drowsiness detection system using Jetson Nano and Raspberry Pi to demonstrate the real-world applicability.

The rest of this paper is organized as follows. Section II summarizes and analyzes the current research work about driving safety. Section III describes the design of our proposed multimodal drowsiness detection system. Section IV shows the performance evaluation of the proposed system. Finally, in Section V, we conclude this paper along with future work.

II. Related Work

For in-vehicle drowsy driver detection, various approaches have been proposed, including computer vision-based methods, physiological signal analysis, and multimodal fusion techniques [7]. Early works primarily focused on single-modal approaches, such as using facial landmarks or eye-tracking systems to assess driver alertness. Recent advancements in deep learning have enabled more sophisticated models that can leverage multiple data sources for improved accuracy.

For computer vision-based methods, many studies have utilized facial landmarks and eye-tracking techniques to detect drowsiness. For example, the MediaPipe Face Mesh model has been widely adopted for real-time facial landmark detection, enabling the calculation of metrics like the Mouth Aspect Ratio (MAR) and Eye Aspect Ratio (EAR) to quantify the degree of mouth opening and eye closure [8]. These metrics have been shown to correlate with drowsiness levels, making them effective indicators for fatigue detection.

For physiological signal analysis, approaches have focused on monitoring heart rate variability (HRV), electroencephalogram (EEG) signals, and other biometric indicators [6]. These signals can provide insights into the driver's mental state and alertness levels. Recent studies have explored the use of wearable devices to continuously monitor these signals, allowing for real-time assessment of drowsiness.

For multimodal fusion techniques, researchers have begun to combine visual and physiological signals to enhance drowsiness detection accuracy. For instance, some studies have integrated facial landmarks with HRV features to create a more comprehensive model that captures both visual and physiological indicators of fatigue [2]. However, many existing systems still rely on traditional LSTM or CNN architectures, which may not fully leverage the potential of attention mechanisms for long-term dependency capture.

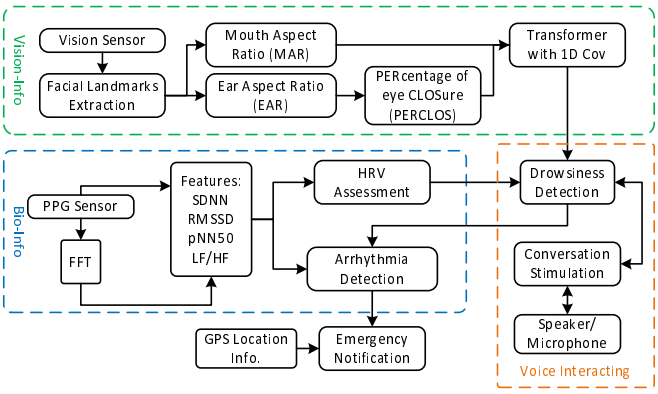


Fig. 2. System architecture of the proposed multimodal drowsiness detection system.

Different from these approaches, our proposed method utilizes a Transformer-based architecture that inherently captures long-range dependencies and relationships between different modalities. This allows for more effective integration of visual and physiological signals, leading to improved drowsiness detection performance.

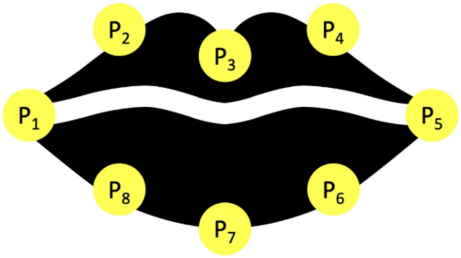
III. System Design

In this section, we describe the design of our proposed on-device multimodal drowsiness detection system. The system is designed to operate in a low-cost embedded environment using Jetson Nano and Raspberry Pi, integrating visual and physiological signals for real-time drowsiness detection. A conceptual system architecture is shown in Fig. 2, which includes the following main components: (1) visual feature extraction using facial landmarks, (2) physiological signal processing using PPG and HRV features, (3) a deep learning-based drowsiness detection model, (4) a conversational interaction module for drowsiness mitigation, and (5) an emergency response function for critical situations.

A. Facial Landmarks

The state of the mouth and eyes of the face was quantitatively analyzed using the MediaPipe Face Mesh model. This model provides a total of 468 face landmarks, of which a subset corresponding to the mouth and eyes was selected to calculate the Mouth Aspect Ratio (MAR) and Eye Aspect Ratio (EAR) that quantify the degree of opening and closing of the mouth and the degree of closing of the eyes.

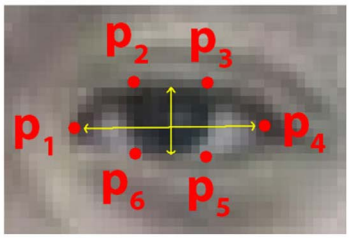
The MediaPipe Face Mesh is a lightweight deep learning-based model that has real-time performance and high accuracy, and can reliably extract landmarks even in various environments. Through this, it was possible to stably collect the



MAR Point

$$MAR = \frac{\| p2 - p8 \| + \| p3 - p7 \| + \| p4 - p6 \|}{3 \| p1 - p6 \|}$$

Fig. 3. Facial landmarks for mouth.



EAR Point

$$\mathrm{EAR} = \frac{\|p_2 - p_6\| + \|p_3 - p_5\|}{2\|p_1 - p_4\|}$$

Fig. 4. Facial landmarks for eyes.

dynamic feature sequence of the mouth and eyes over time from camera images.

MAR was calculated to extract the features of the mouth, which is defined as the average distance between the top and bottom landmarks of the mouth divided by the distance between the left and right landmarks of the mouth, as shown in Fig. 3. MAR is used to quantify the degree to which the mouth is opened in real time. Since the mouth structure varies from person to person, the baseline MAR was set for each individual by collecting data in the state of mouth opening for a certain period of time, and the relative MAR increase rate was calculated based on this to determine whether yawning was present. The frequency of yawning within a certain period of time, for example, the number of yawns per 5 minutes, was used as the main index for determining fatigue rather than single yawning. Yawning duration was excluded from the analysis due to large individual differences and low discrimination power.

The degree of eye closure was measured through EAR, which is defined as the ratio of the average vertical length of the eye divided by the horizontal length. Fig. 4 shows the landmarks used to calculate EAR. If the EAR falls below a specific threshold, it is judged that the eyes are closed. In addition, PERCLOS, the rate at which the eyes were closed for a certain period of time, and the duration of the closing between the eyes and the reopening were extracted together and used as fatigue indicators. Excluding short blinkers, only meaningful closing was reflected in the analysis.

Since the features extracted from the mouth and eyes have a time series shape that changes over time, a model structure that combines a one-dimensional convolutional neural network (CNN), a recurrent neural network (e.g., LSTM), and an attention mechanism was designed to deal with this. For each structure, a one-dimensional CNN was applied to extract characteristic patterns and reduce noise, and the extracted feature maps were combined to form one integrated feature vector. This integrated sequence was then entered into the LSTM layer to learn the gradual accumulation of time dependence and fatigue, and the attention mechanism allowed high weights to be given at critical points for sleepiness judgment.

All MAR and EAR values were normalized and used as rates of change relative to individual reference values, not absolute values. Through this, it was possible to more accurately reflect changes in patterns without being affected by individual physiological differences or basic conditions.

Finally, a model for binary classification of drowsiness was constructed based on the LSTM and the output of the attention structure. A cross entropy loss function was used for learning, and model performance was evaluated through accuracy of the detection.

B. PPG and HRV Features

In this study, for the detection of drowsiness based on physiological signals, ECG data of the DROZY dataset [9] were used to extract HRV (Heart Rate Variability) features that are used for model learning. HRV is a representative biomarker that reflects the activity state of the autonomic nervous system and is known to be effective in quantitatively tracking physiological changes such as drowsiness.

Since the DROZY dataset provides ECG-based RR interval information [10], the several HRV features were calculated based on the information, as shown in Table I. HRV generally includes various analysis indicators divided into a time domain, a frequency domain, and a nonlinear domain, and in this study, time and frequency domain features were mainly used.

First, in the time domain analysis, features capable of grasping the overall variability and short-term change in heart rate intervals were extracted. Typically, SDNN represents the standard deviation of the entire NN interval (normal to normal interval) and represents the overall variability of the autonomic nervous system. RMSSD refers to the square root of the mean difference squared between two consecutive heart rate intervals, which is mainly interpreted as a measure of parasympathetic nervous activity. NN50 and pNN50 represent the number and overall ratio of cases where the difference between the two RR intervals is more than 50 ms, and also reflects the effect of the parasympathetic nervous system.

TABLE I
SUMMARY OF THE MAIN HRV FEATURES EXTRACTED FROM THE DROZY DATASET

Feature	Description	Time Domain	Frequency Domain
SDNN	Standard deviation of NN intervals	✓	
RMSSD	Root mean square of successive differences	✓	
NN50	Number of pairs of successive RR intervals differing by more than 50 ms	✓	
pNN50	Proportion of NN50 to total number of RR intervals	✓	
LF	Low frequency power (0.04-0.15 Hz)		✓
HF	High frequency power (0.15-0.4 Hz)		✓
LF/HF	Ratio of LF to HF power		✓

In the frequency domain analysis, the time series data of the RR interval was converted into a power spectrum through a frequency analysis (such as a Fast Fourier Transform or Welch method) and the power value of a specific frequency band was calculated. Among them, low frequency (LF) (0.04-0.15 Hz) reflects the complex reaction of sympathetic and parasympathetic nerves, and high frequency (HF) (0.15-0.4 Hz) mainly represents the activity of parasympathetic nerves. The LF/HF ratio is an index that quantifies the balance state of the autonomic nervous system and tends to show a clear pattern of changes according to the drowsiness state.

The window configuration for HRV feature extraction is as follows. A fixed window of 5 minutes is generally used in standard HRV analysis, but in this study, the entire session for a total of 9 to 10 minutes was divided using sliding windows at regular intervals (e.g., 5 minutes window of a 30-second movement interval) to match the composition and consistency of DROZY data, and HRV features were calculated for each section. As a result, it was possible to construct a sequence of feature changes over time, which was used as the basis data for entering the time series learning model.

In the signal preprocessing step, noise removal and outlier removal were preceded to ensure accurate RR interval extraction from the ECG signal. In particular, adaptive filtering and range-based outlier removal algorithms were applied to detect noise, errors, and remove artifacts included in the signal. Since the quality of the RR interval directly affects the accuracy of HRV analysis, the corresponding preprocessing process was regarded as an essential step to secure the reliability of HRV features.

The HRV features derived in this way are integrated into multimodal input along with visual features and used as physiological signal-based input of the fatigue state prediction model. Physiological signals provide information complementary to visual signals, so that improved drowsiness detection performance can be expected compared to a single modality-based model.

C. Conversations Stimulation

Fatigue or drowsiness that occurs while driving causes attention loss, reaction time delay, cognitive processing ability, etc., and is pointed out as one of the main causes of traffic accidents. In particular, in a monotonous and repetitive driving environment, the level of arousal of the driver gradually decreases, thereby increasing the risk of an accident. Thus, the proposed system provides a drowsiness mitigation function that presents a simple conversation and language game based on natural language so that a driver may recover his or her arousal state without external stimulation after detecting such a drowsiness state in real time.

Related studies have shown that conversations with digital assistants while driving are effective in relieving drivers' drowsiness and inducing cognitive arousal. A study analyzed that when a driver had a conversation with a digital interactive agent, they showed better lane keeping, faster response to dangerous situations, and increased distraction range, which was associated with an overall improvement in attention span. The study also reported that drivers' levels of drowsiness during conversations decreased, and their efforts to stay awake were also reduced.

In a similar context, another study [11] confirmed that cognitive stimulation through language-based games or conversations significantly reduced drivers' fatigue and sleepiness indicators in a 20-30-minute fatigue state. In particular, the study experimentally demonstrated that simple conversation and response-inducing questions alone can achieve fatigue recovery and attention-switching effects, suggesting its potential for use in autonomous driving technology and driver assistance systems.

Therefore, the drowsiness mode processing function of this system is a way to provide active cognitive stimulation beyond simple warning sounds, and only short interactions can contribute to relieving drowsiness and restoring the driver's cognitive arousal.

D. Emergency Response Function

When the driver's condition exceeds a certain standard and deteriorates, a simple drowsy warning alone cannot prevent the accident. To compensate for this, the system determines the user's condition as an emergency situation and performs an immediate response procedure when a certain condition is satisfied.

The emergency assessment condition is set when, for example, drowsiness persists for a certain period of time (e.g., 1 minute) or when a drowsiness pattern is repeatedly detected. These criteria refer to the fatigue scoring index used in other drowsiness detection systems, and the threshold value considered an emergency if there is no response within a certain period of time [12]. When an emergency assessment is made, the system presents a simple voice or text-based question (e.g., dizziness, chest pain) to the user, attempts to

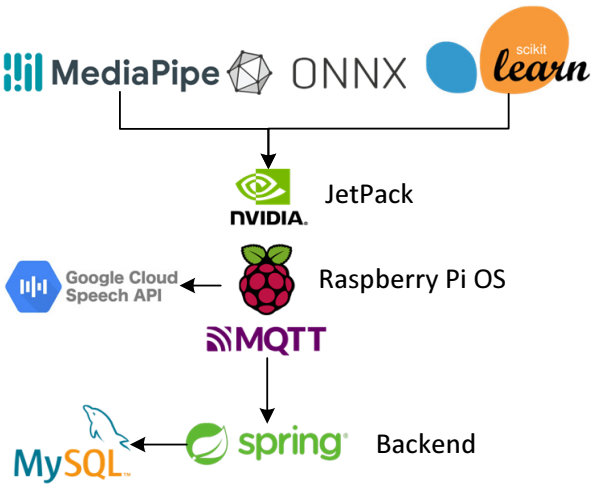


Fig. 5. Software stack of the proposed system.

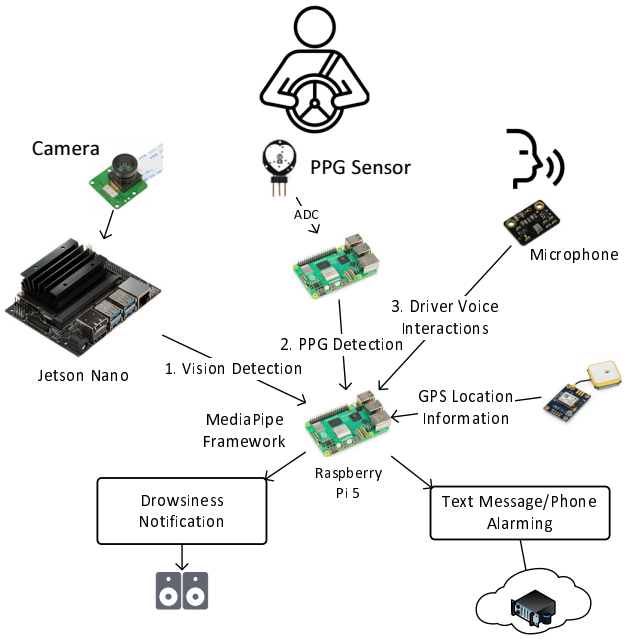


Fig. 6. The implementation of the proposed multimodal drowsiness detection system.

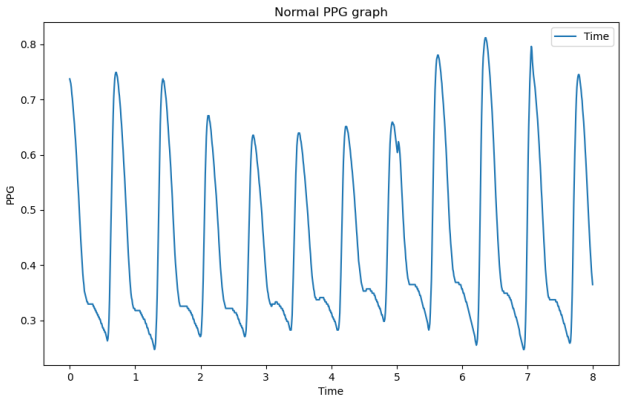


Fig. 7. The normal PPG signals in the dataset.

check symptoms, and at the same time obtains the vehicle's location information through GPS.

Thereafter, the collected user response and location information are transmitted to the server and recorded, and are transmitted in real time to a guardian or acquaintance registered in advance by the user. In addition, when there is no response from the user or when it is determined that the symptom is urgent, the report is automatically made to an emergency medical institution (e.g., 119 or 911). This automation structure is an important function for securing golden time for emergency response, and is being applied in a similar form in systems for elderly drivers, and the importance of early intervention is repeatedly emphasized in the medical community.

If the user's consciousness is restored, the system automatically returns to the normal operation mode, otherwise repeatedly checking the user's condition and continuing monitoring until emergency rescue is made.

IV. Performance Evaluation

In this study, various experiments were conducted to verify the performance of the drowsiness detection model using visual information and physiological signals. Fig. 5 shows the software stack, and Fig. 6 shows the overall implementation of the proposed multimodal drowsiness detection system.

First of all, the data used in the experiment consist of the DROZY dataset, which includes facial landmarks extracted from camera images, especially coordinate information of the mouth and eyes, and PPG or ECG-based HRV signals containing heart rate variability information. The label for drowsiness was based on annotations provided in the dataset, and in some cases, RMSSD, LF/HF ratio, or EEG-based indicators derived from HRV were used as labels.

The feature extraction process was divided into two modalities: visual information and physiological signals. In visual information, after extracting face landmarks by applying MediaPipe Face Mesh, MAR and EAR values were obtained through the calculation of ratios to the mouth and eyes. In addition, fatigue-related indicators such as the number of yawns, the number of eye closures, and PERCLOS were quantified. Since the dataset is pre-recorded, there was a limitation that it was difficult to measure the baseline for each individual user. To compensate for this, a correction procedure was performed for each user using the average value of the section classified as normal or through z-score normalization.

In the physiological signal modality, time domain and frequency domain features such as RMSSD, LF, HF, LF/HF ratio, and SDNN were extracted from HRV. Table. I shows the extracted HRV features. Fig. 7 shows the normal PPG signals from the dataset, and Fig. 8 shows the real-time PPG signal collection and the testing environment, which uses a analog-digital converter with the SPI interface to forward the signal to the Raspberry Pi 5 board. These features were generated for each section through a sliding window of a certain length,

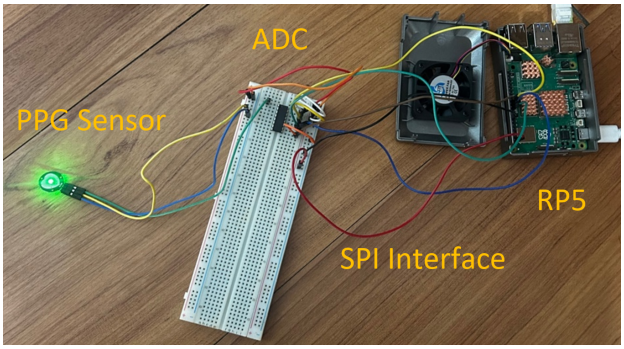


Fig. 8. The real-time PPG signal collection and testing environment.

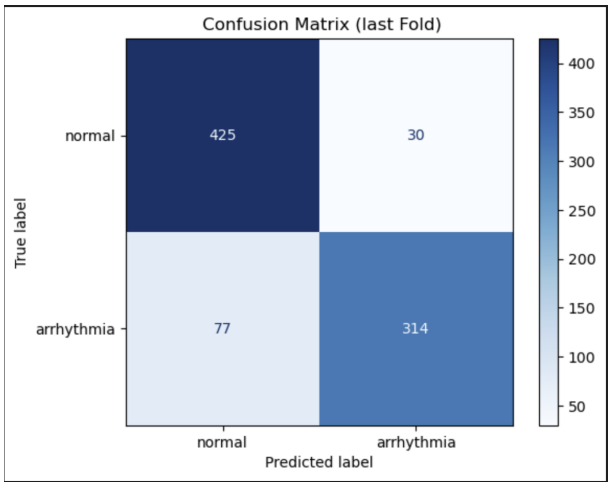


Fig. 9. The confusion matrix for detecting arrhythmia from PPG signals.

and were generally based on a section of 30 seconds.

The model structure consisted of a multi-step processing process. First, visual features, i.e., MAR, EAR, and related figures, were processed through a one-dimensional (1-D) CNN, and HRV features were also input to a separate 1-D CNN to extract characteristics. After that, the feature vectors obtained from the two modalities were combined into one, and the temporal dependence was learned by inputting them into the LSTM layer. In the last step, the attention mechanism was applied to give weight to the point of time when the contribution to determining whether or not to sleepiness was high, and binary classification was finally performed.

In the experimental process, the dataset was divided into training, verification, and test sets, and a K-fold or subject-wise split was applied for cross-validation. In addition, the classification accuracy for each class was visually analyzed through the confusion matrix, as shown in Fig. 9.

A comparative experiment was also conducted in parallel. Through this, the performance difference among different models was evaluated. Table II shows the performance of the transformer-only, the LSTM-only structure, and the Transformer with 1D convolution model. From the results, it was confirmed that the Transformer model showed better performance than the LSTM model, and when 1D convolution was added in front of the Transformer, the performance was improved significantly up to 79.35% of the accuracy to recognize drowsiness.

TABLE II
PERFORMANCE COMPARISON OF DIFFERENT MODELS

Model	Accuracy
Transformer-only	59.98%
LSTM-only	53.04%
Transformer + 1D Convolution	79.35%

V. CONCLUSION

In this paper, we presented the feasibility of a driver condition monitoring system that effectively integrates multimodal data (visual + physiological signal) using low-cost embedded hardware and secures high accuracy performance by applying a Transformer-based model. One limitation of the current study is that the performance was evaluated partially on the Jetson Nano platform due to hardware constraints, and the real-time performance in actual driving environments was not fully verified. Future research will further check the real-time performance and increase the reliability of the system by securing empirical data in actual vehicle environments.

REFERENCES

[1] H. H. Jeong, Y. C. Shen, J. P. Jeong, and T. T. Oh, "A comprehensive survey on vehicular networking for safe and efficient driving in smart transportation: A focus on systems, protocols, and applications," *Vehicular Communications*, vol. 31, p. 100349, 2021.

[2] C. Kodikara, S. Wijekoon, and L. Meegahapola, "Fatiguesense: Multi-device and multimodal wearable sensing for detecting mental fatigue," *ACM Trans. Comput. Healthcare*, vol. 6, no. 2, pp. 1–36, 2025.

[3] H. Pan, S. Tong, X. Wei, and B. Teng, "Fatigue state recognition system for miners based on a multimodal feature extraction and fusion framework," *IEEE Trans. Cogn. Dev. Syst.*, vol. 17, no. 2, pp. 410–420, 2025.

[4] S. Yu, Q. Yang, J. Wang, and C. Wu, "Fedusl: A federated annotation method for driving fatigue detection based on multimodal sensing data," *ACM Trans. Sen. Netw.*, Apr. 2024.

[5] L. Yu, X. Yang, H. Wei, J. Liu, and B. Li, "Driver fatigue detection using ppg signal, facial features, head postures with an lstm model," *Heliyon*, vol. 10, no. 21, p. e39479, 2024.

[6] G. Du, L. Zhang, K. Su, X. Wang, S. Teng, and P. X. Liu, "A multimodal fusion fatigue driving detection method based on heart rate and perclos," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 11, pp. 21 810–21 820, 2022.

[7] S. Cao, P. Feng, W. Kang, Z. Chen, and B. Wang, "Optimized driver fatigue detection method using multimodal neural networks," *Scientific Reports*, vol. 15, no. 1, p. 12240, 2025.

[8] Y. Kartynnik, A. Ablavatski, I. Grishchenko, and M. Grundmann, "Real-time facial surface geometry from monocular video on mobile gpus," 2019.

[9] 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*, 2016, pp. 1–7.

[10] K. Fujiwara, H. Iwamoto, K. Hori, and M. Kano, "Driver drowsiness detection using r-r interval of electrocardiogram and self-attention autoencoder," *IEEE Trans. Intell. Veh.*, vol. 9, no. 1, pp. 2956–2965, 2024.

[11] D. R. Large, G. Burnett, V. Antrobus, and L. Skrypchuk, "Driven to discussion: engaging drivers in conversation with a digital assistant as a countermeasure to passive task-related fatigue," *IET Intell. Transp. Syst.*, vol. 12, no. 6, pp. 420–426, 2018.

[12] N. H. T. S. A. (NHTSA), "Drowsy driving and automobile crashes," 2016. [Online]. Available: https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/812115