Driver State Classification via Multimodal Physiological Signal Fusion and Hybrid Decomposition with 1D CNN-BiGRU Network

He Liu, IAENG, Member, JinTao Yu, TingWei Liang

***Abstract*—Driver state detection is essential for ensuring traffic safety; however, current research is hindered by limited generalization across datasets, inadequate exploration of the spatio-temporal feature coupling inherent in multimodal physiological signals, and model bias stemming from category imbalance. This study introduces an innovative hybrid framework that integrates graph signal processing, empirical mode decomposition (EEMD), and topological decomposition for feature extraction. Additionally, it incorporates generative adversarial network (GAN) data augmentation within a 1D CNN-BiGRU multimodal fusion architecture, which has been validated across four diverse datasets. Experimental results indicate that the average accuracies of individual signals optimized through feature decomposition achieve 75.5% and 76.7% using traditional models (Decision Trees/K-Nearest Neighbors), while these figures increase to 79.5% and 81.7% within the 1D CNN-BiGRU network. Notably, the proposed 1D CNN-BiGRU model attains an impressive accuracy of 98.1% on the Driver Stress dataset (PSY) following multimodal fusion, significantly surpassing the performance of comparative models (CNN: 90-94%, BiGRU: 94%, Transformer: 97%, TCN: 95%) and improving by 16% over the best unimodal results. Furthermore, the model exhibits remarkable cross-domain generalization capabilities, maintaining high accuracy levels exceeding 96% on non-driving scenario datasets (CASE: 98%, DEAP: 96%, EmoWear: 97%). Through the implementation of global-local feature decoupling and an adversarial data balancing mechanism, this study establishes a novel paradigm characterized by both high robustness and strong interpretability, effectively addressing the challenges associated with multi-scene driver state detection.**

***Index Terms*—Driving Behaviour Detection; Graph Signal Processing; EEMD Decomposition; Topological Decomposition; 1D Convolutional Neural Network (1D CNN); Bidirectional Gating Unit (BIGRU).**

I. INTRODUCTION

The real-time classification of driver behavioral states poses a significant challenge in the fields of intelligent transportation and driver safety. Accurately identifying driver states—such as fatigue, distraction, and mood fluctuations—is essential for implementing active safety systems and has the potential to substantially reduce the risk of traffic accidents [1], [2]. Although existing research has demonstrated the effectiveness of physiological signals—including electrodermal activity (EDA), electrocardiogram (ECG), blood volume pulse (BVP), and skin temperature (SKT)—in monitoring driver states, there is still a need for enhancements in the efficiency of deep feature extraction and multimodal fusion techniques. Therefore, developing a robust, low-latency classification system through advanced signal processing and optimized model design is an urgent and critical challenge.

In recent years, researchers both nationally and internationally have engaged in comprehensive investigations into the recognition of driver states. Xiang et al.[3] developed a multi-modal dataset (MMDE) that combines facial expressions and physiological signals, demonstrating an enhancement in emotion recognition accuracy of 11.28%. However, their study did not provide a thorough examination of the spatiotemporal topological characteristics of physiological signals. Lalwani et al.[4] introduced a CNN-BiLSTM-BiGRU model for analyzing wearable sensor data, achieving an impressive accuracy of 99.33% in human activity recognition; nonetheless, their framework was not tailored to the intricate temporal features inherent in driving scenarios. Zaki et al.[5] proposed a hybrid deep learning model, CBGA, which integrates CNN, bidirectional gated recurrent units, and attention mechanisms, attaining 97.75% accuracy in multi-label classification for sentiment analysis. However, this model primarily focused on text classification and did not incorporate physiological signal processing. Ma et al.[6] presented a method utilizing Transformers and pseudo-label multi-task learning within a digital twin framework (DDT), which significantly enhanced the recognition of distracted behaviors, yet did not fully leverage the multi-scale information available from physiological signals. Wei et al.[7] introduced an explainable recommendation algorithm based on deep learning, employing Bi-LSTM and MCNN to extract multi-dimensional features, resulting in a 1.57% improvement in accuracy; however, their approach is predominantly applicable to recommendation systems. Chen et al.[8] and Ying et al.[9] demonstrated the benefits of multimodal methodologies through the use of EEG graph neural networks and audio-visual fusion models, respectively. Nonetheless, the former relies on specialized equipment, while the latter fails to account for the dynamic correlations of physiological signals. Hu et al.[10] proposed a hybrid model that combines CNN and Transformer architectures to extract multi-scale spatiotemporal features from EEG signals, achieving an accuracy of 91.26%

Manuscript received May 20, 2025; revised August 7, 2025.

This work was supported by the National Natural Science Foundation of China (Grant Nos. 12422213, 12372008); the National Key R&D Program of China (Grant No. 2023YFE0125900); the Natural Science Foundation of Heilongjiang Province (Grant No. YQ2022A008); and the Basic Research Programs of Heilongjiang Provincial Universities (Grant No. 2023KYYWF0980).

He Liu is a postgraduate student at Harbin University of Commerce in Harbin, China (Email: lh15537566063@163.com).

JinTao Yu is a professor at Harbin University of Commerce in Harbin, China (Corresponding author to provide email: 101129@hrbcu.edu.cn).

Tingwei Liang is a professor at the Harbin Institute of Technology in Harbin, China (Email: liangtingwei@hit.edu.cn).

in four-category emotion classification; however, this is restricted to the single modality of EEG. Hong et al.[11] attained an identification accuracy exceeding 86% through an anomaly detection and multispectral imaging-based facial expression recognition model (AF-MSI), but their focus was primarily on facial visual features, lacking integration of physiological signal data. In conclusion, existing research continues to exhibit limitations concerning deep signal representation, multimodal collaboration, and adaptability to driving scenarios.

In recent years, novel methodologies such as graph signal processing (GSP) [12], topological decomposition, and ensemble empirical modal decomposition (EEMD) have emerged, providing advanced strategies for analyzing physiological signals [13]. GSP is particularly effective in capturing the spatial and topological relationships inherent in physiological signals. When integrated with EEMD, it enables the adaptive extraction of multiscale features from nonstationary and non-smooth signals. Furthermore, the 1D CNN-BiGRU network synergizes localized feature extraction with long-term dependency modeling, thereby establishing a robust framework for multimodal fusion. The combination of these techniques has the potential to overcome the limitations of conventional methods regarding feature representation and model generalization. This study presents a novel framework in which Graph Signal Processing (GSP), combined with topological decomposition, constructs a dynamic graph representation of physiological signals. Additionally, Empirical Mode Decomposition (EEMD) is employed to extract their multiscale intrinsic mode components. Subsequently, a one-dimensional Convolutional Neural Network-Bidirectional Gated Recurrent Unit (1D CNN-BiGRU) network is utilized to integrate spatiotemporal features, facilitating precise classification of driver states. Experimental findings indicate that the proposed approach significantly outperforms existing models across four public datasets, achieving an improvement in classification accuracy ranging from 6.5% to 12.3%. By elucidating the intrinsic correlations between physiological signals and driving behavior, this research contributes to the theoretical advancement of multimodal fusion technology and provides substantial support for real-time driving safety warning systems, with significant implications for engineering applications.

II. PROPOSED METHOD

This section presents a model for monitoring driver conditions that employs a multimodal fusion of spatiotemporal features derived from four distinct physiological signals.

A. Driver Condition Monitoring Model Architecture

The comprehensive architecture of the driver state monitoring model is illustrated in Figure 1. This proposed model consists of two primary components:

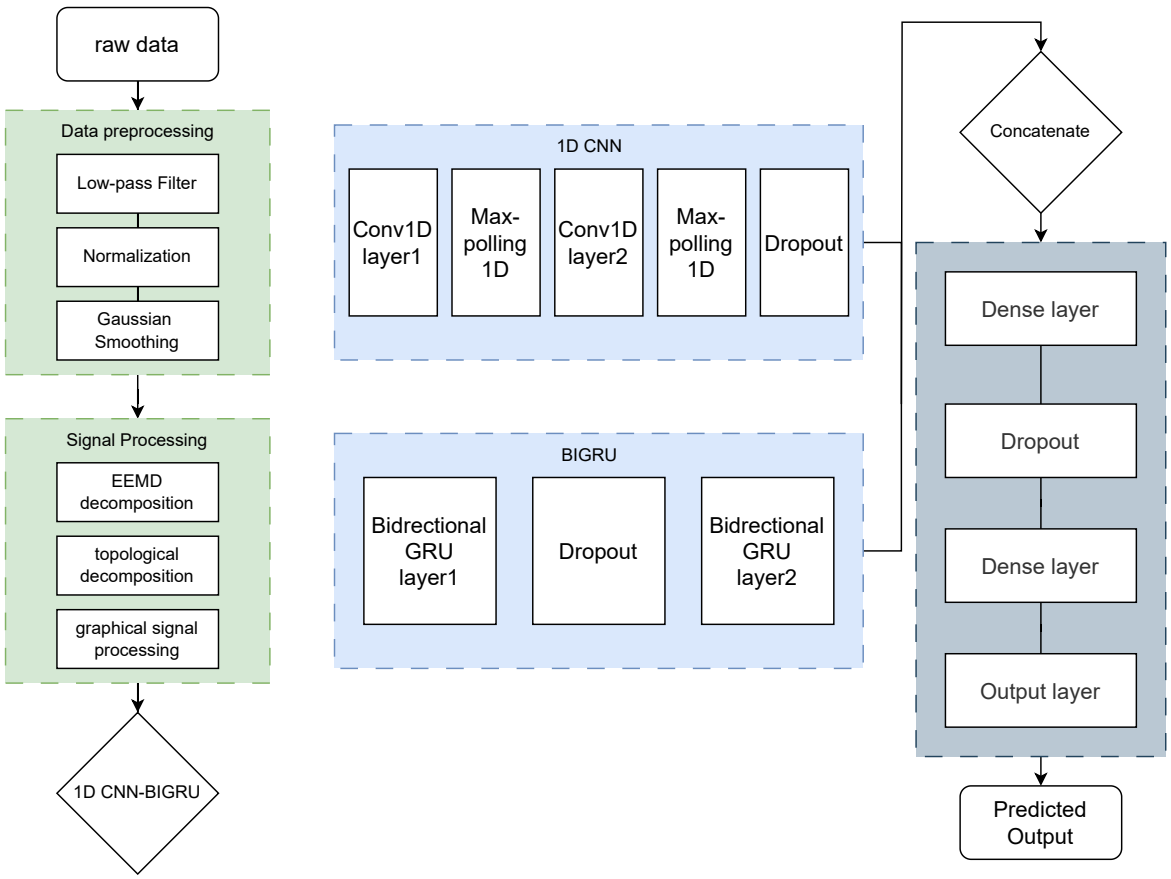

Fig. 1. Driver Condition Monitoring Model.

an advanced spatiotemporal feature extraction module and a one-dimensional Convolutional Neural Network-Bidirectional Gated Recurrent Unit (1D CNN-BiGRU) network. The initial module surpasses conventional neural networks in accurately extracting spatiotemporal features from the driver's physiological signals, while the latter is responsible for learning and classifying various driver states.

B. Signal Pre-Processing

In the field of physiological signal processing, four distinct types of signals—namely, electrodermal activity (EDA), electrocardiographic signals (ECG), blood volume pulse (BVP), and skin temperature (SKT)—exhibit significant commonalities. These signals are derived from physiological processes and encapsulate the intricate dynamics of the autonomic nervous system. Furthermore, they are particularly susceptible to noise and disturbances arising from both physiological and environmental factors. Given that these signals are typically acquired from complex physiological systems, the raw data often contain various forms of unwanted interference, including power-line noise, baseline drift, physiological artifacts, and instrumental noise. To address these challenges, we have developed a comprehensive preprocessing pipeline.

Step 1: Data Quality Assessment. The integrity and duration of the signal are rigorously evaluated to eliminate samples that do not meet established minimum criteria, thereby ensuring the reliability of subsequent analyses. Samples containing fewer than a specified number of data points (e.g., 20) are discarded to minimize the impact of low-quality data on the overall analysis.

Step 2: Outlier Identification and Removal. Specialized detection algorithms are employed to identify and eliminate extreme values from the signals. This process effectively removes isolated noise caused by sensor malfunctions, human interference, or random variations, thereby significantly enhancing the stability of the signal.

Step 3: Correction of Baseline Drift. A technique for removing drift is implemented to eliminate low-frequency oscillations and gradual signal drifts. This step is essential for preserving critical signal characteristics and mitigating long-term fluctuations caused by sensor bias, variations in skin resistance, and other influencing factors.

Step 4: Noise Reduction and Signal Enhancement. Initially, a low-pass filter is applied to reduce high-frequency noise. Subsequently, signal normalization is performed using zero-mean and unit-variance transformations to eliminate discrepancies in scale between signals. Further smoothing is achieved through Gaussian filtering, which mitigates minor fluctuations while preserving essential signal features. Finally, the signals are resampled to a standardized length, ensuring consistency and reliability for feature extraction and machine learning applications.

Step 5: Data Balancing. To address the issue of class imbalance present in physiological signal datasets, we implement a data-balancing strategy utilizing generative adversarial networks (GANs). First, we determine the target balanced sample size by dividing the total sample size by the number of classes. Next, we apply random undersampling to the majority class and develop a GAN model specifically for the minority class to augment the dataset. The GAN architecture consists of a generator and a discriminator; the generator is responsible for producing realistic physiological signal samples from a 128-dimensional latent space, while the discriminator's role is to distinguish between authentic and generated samples. During the training process, we incorporate label smoothing techniques, which involve adding noise to real labels (0.9 ± 0.1) to reduce the risk of overfitting. Additionally, we implement a dynamic stopping mechanism, whereby training is automatically halted when the discriminator's accuracy exceeds 95% or when the disparity between the losses of the two networks falls below 0.01, thereby preventing model collapse. The synthetic samples generated are subsequently restored to their original scale through inverse normalization and are randomly assigned corresponding subject labels to ensure biological validity. This procedure results in a complete balance across all emotional categories, thereby establishing a robust foundation for the training of subsequent classification models.

This comprehensive preprocessing technique significantly enhances the signal-to-noise ratio, resulting in a more reliable dataset for the subsequent analysis and interpretation of physiological signals.

C. Spatio-Temporal Feature Extraction Utilizing Graph Signal Processing, Topological Decomposition, and Empirical Modal Decomposition.

In the field of physiological signal processing, traditional neural network methods for feature extraction often fail to sufficiently capture the complex intrinsic structures and dynamic characteristics inherent in multimodal physiological signals. This research introduces an innovative framework that integrates graph signal processing, topological decomposition, and empirical modal decomposition techniques to enhance the depth and accuracy of signal feature extraction. This improvement is accomplished through the multidimensional characterization and fusion of features.

Conventional methods for extracting features from physiological signals exhibit significant limitations when addressing non-stationary signals. These limitations are primarily evident in the incomplete extraction of features within the time-frequency domain and the inability to model correlations between different signal modalities. The proposed approach, which integrates graph signal processing with Ensemble Empirical Mode Decomposition (EEMD), effectively addresses these challenges through a multi-scale spatio-temporal decoupling strategy. Specifically, EEMD employs adaptive decomposition techniques to break down complex non-stationary physiological signals into multiple intrinsic mode functions (IMFs), with each IMF representing signal components across various time scales. This decomposition is independent of predefined basis functions, allowing for the adaptive extraction of intrinsic temporal patterns within the signal, effectively distinguishing noise from useful information, and mitigating the frequency leakage issues associated with traditional Fourier transforms in the context of non-stationary signals.

Concurrently, graph signal processing constructs a graph network structure for multi-modal physiological signals, utilizing graph Laplacian operators and graph filters to extract topological features in the spatial domain, thereby capturing cross-modal association patterns. The synergistic interaction of these two methodologies facilitates a comprehensive integration of spatio-temporal features: the IMF components derived from EEMD are processed within the graph domain, enabling each time scale of the signal to leverage spatial topological information for enhanced feature extraction. This collaborative mechanism not only preserves the temporal dynamics of the signal but also fully capitalizes on the spatial correlations among multimodal signals, resulting in richer and more discriminative feature representations and significantly improving the representation capabilities of non-stationary physiological signals.

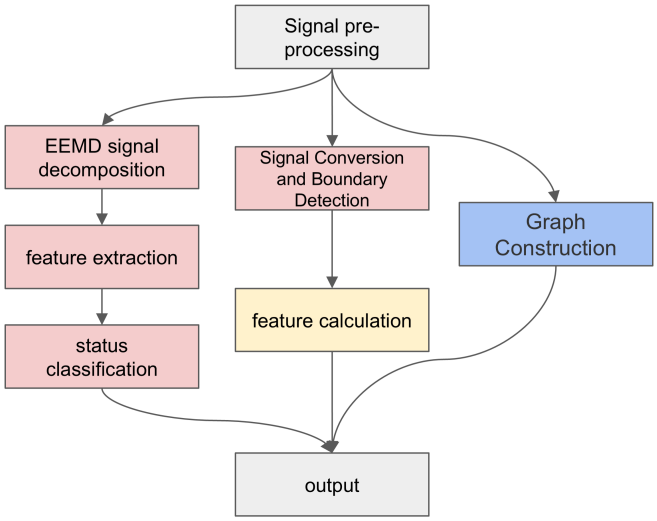

Fig. 2. Spatio-Temporal Feature Extraction Module.

The spatiotemporal feature extraction module, as illustrated in Figure 2, consists of four components: graph feature extraction, topological feature extraction, EEMD feature extraction, and the layers for feature selection and integration. The procedural framework for developing the spatiotemporal feature extraction module, which is based on these methodologies, is outlined as follows:

Step 1: The initial phase involves utilizing the K-nearest neighbor graph construction technique to derive a topological structure graph from the signal. This process begins with downsampling and smoothing the original signal using a sliding window approach, resulting in a filtered signal denoted as $s(t)$. Subsequently, the signal is quantized with a specified quantization step size Q, expressed mathematically as

$$s_q(t) = round(\frac{s(t)}{Q}) \cdot Q \tag{1}$$

The vertex set V is established based on the unique values obtained after quantization. The K-nearest neighbor algorithm is employed to calculate the Euclidean distances between the vertices, defined as follows:

$$d_{ij} = \sqrt{\sum_{k=1}^{n}(x_{ik} - x_{jk})^2} \tag{2}$$

Thereby, the edge set E is constructed, forming an undirected graph $G = (V, E)$. Multiple global features are extracted from the graph, including load centrality, which is calculated as

$$C_L(v) = \sum_{s,t \in V} \frac{\sigma(s,t|v)}{\sigma(s,t)} \tag{3}$$

In this context, $\sigma(s,t)$ denotes the total number of shortest paths between vertices s and t, while $\sigma(s,t|v)$ represents the number of shortest paths that pass through vertex v. Additional features include harmonic centrality, group count, graph diameter, and graph radius. The connectivity analysis of the graph improves the global characteristics of the feature description, ensuring the connectivity of each graph. The formulas for harmonic centrality, graph diameter, and graph radius are expressed as follows:

$$C_H(v) = \sum_{t \in V \setminus v} \frac{1}{d(v,t)} \tag{4}$$

$$D(G) = \max_{u,v \ inV} d(u,v) \tag{5}$$

$$R(G) = \min_{v \in V} \max_{u \in V} d(u,v) \tag{6}$$

Step 2: The second phase involves decomposing the signal using a customized Ensemble Empirical Mode Decomposition (EEMD) algorithm to extract intrinsic mode functions (IMFs). Statistical features are then derived from each IMF, including fundamental statistical metrics such as mean, variance, peak count, and amplitude, defined as follows:

$$A = \max(IMF) - \min(IMF) \tag{7}$$

Furthermore, three critical Hjorth parameters are computed: activity, which quantifies signal energy; mobility, which indicates variations in average frequency; and complexity, which describes the spectral characteristics of the signal. The relevant formulas are as follows:

$$Activity = \sigma^2 = var(IMF) \tag{8}$$

$$Mobility = \sqrt{\frac{var(dIMF)}{var(IMF)}} \tag{9}$$

In instances where the Information Measure of Complexity (IMF) is positive, entropy is also employed to assess the uncertainty of the signal. Collectively, these features encapsulate the diversity of the signal within the time domain, with each feature reflecting distinct attributes of the signal that are subsequently utilized for emotion

classification tasks. The formulas for complexity and entropy are presented as follows:

$$Complexity = \frac{Mobility(dIMF)}{Mobility(IMF)} \tag{10}$$

$$Entropy = -\sum(IMF \cdot \log(IMF)) \tag{11}$$

Step 3: The final phase involves applying the discrete Fourier transform to map physiological signals onto the complex plane. The geometric boundary representation of the signal is constructed using the Alpha Shape algorithm, which facilitates the extraction of geometric features such as the area of the convex hull and the perimeter of the boundary. Feature standardization is performed using the MinMaxScaler, and the valence and arousal parameters are integrated for the classification of emotional states. This integration enables multi-dimensional emotional recognition that encompasses categories ranging from neutral to pleasant, bored, relaxed, and fearful. The formulas for the discrete Fourier transform and feature standardization are outlined as follows:

$$X[k] = \sum_{n=0}^{N-1} x[n] e^{-j2\pi kn/N} \tag{12}$$

$$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{13}$$

Step 4: Implement feature selection algorithms, such as the Anderson-Darling normality test, ReliefF, or Recursive Feature Elimination (RFE), on the graph, features to identify the most relevant dimensions for the classification task. This process reduces feature redundancy, thereby enhancing computational efficiency and model generalization while preserving essential information.

Step 5: The refined graph features, along with temporal features derived from deep learning techniques and statistical features obtained through Empirical Mode Decomposition (EEMD), are integrated to create a cohesive temporal-spatial feature vector. This vector encapsulates both temporal and spatial information, providing a comprehensive representation of features for the subsequent classification model.

By employing multi-scale and multi-dimensional feature fusion, this approach preserves the intrinsic characteristics of each modal signal while uncovering complex signal patterns that conventional methods often overlook. This leads to a more nuanced and discriminative characterization of features, thereby enhancing the recognition and prediction of physiological states.

D. Driver Monitoring Model Based on 1D CNN-BIGRU

The 1D CNN-BIGRU network module, as illustrated in Figure 3, consists of a one-dimensional convolutional neural network (CNN) layer, a bidirectional gated recurrent unit (BIGRU) layer, and a feature fusion layer. The procedure for constructing this network is outlined as follows:

Step 1: Specify the input and output variables of the network. The input variables consist of feature-extracted

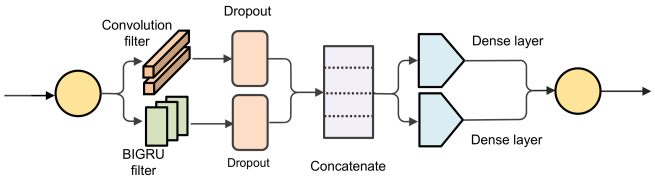

Fig. 3. 1D Network Module Diagram of CNN-BIGRU.

signals from electrodermal activity (EDA), electrocardiogram (ECG), skin temperature (SKT), and blood volume pulse (BVP), each represented as distinct dimensions. The output variables are multiclass classification labels that correspond to various state categories. The input shape of the network aligns with the dimensions of the features, while the output reflects the classification results. The input is denoted as $X = \{X_{EDA}, X_{ECG}, X_{SKT}, X_{BVP}\}$, and the output is represented as $Y \in \{1, 2, \ldots, K\}$.

Step 2: Utilize the StandardScaler method to normalize the feature data. Each individual signal feature set is normalized independently to achieve a mean of 0 and a standard deviation of 1, thereby mitigating the disparity in magnitude among the features. This normalization process enhances the stability of the model and accelerates the speed of convergence. The formula for normalization is as follows:

$$X_{normalized} = \frac{X - \mu}{\sigma} \tag{14}$$

Step 3: Partition the normalized dataset into a training set and a test set. The training set is used for model fitting, while the test set is employed to evaluate model performance. The train-test split methodology is implemented, allocating 80% of the data to the training set and 20% to the test set. This approach is designed to maintain the consistency of label distribution, thereby enhancing the model's generalization performance.

$$\mid X_{train} \mid = 0.8 \mid X \mid, \mid X_{test} \mid = 0.2 \mid X \mid \tag{15}$$

Step 4: Develop the specifications and processing architecture of the neural network. Construct independent subnetwork modules for each input signal, which should include convolutional processing for Electrodermal Activity (EDA) and Skin Temperature (SKT), as well as bidirectional Gated Recurrent Unit (GRU) processing for Electrocardiogram (ECG) and Blood Volume Pulse (BVP). After aligning the features from the subnetworks, perform high-level fusion through a fully connected layer. Finally, incorporate a Softmax classification layer to generate the prediction outcomes. The equations governing feature fusion and the final output are represented as follows:

$$F = Concat(F_{EDA}, F_{ECG}, F_{SKT}, F_{BVP}) \tag{16}$$

$$P(y|X) = Softmax(WF + b) \tag{17}$$

Step 5: Finalize the training of the network and the output of results. Utilize the cross-entropy loss function in conjunction with the Adam optimizer for model

training, while establishing a validation set to assess model performance. Upon completion of the training phase, evaluate the model using the test set, compute the classification accuracy, and generate a classification report that includes precision, recall, and the F1 score. The formulas for the loss function and accuracy are outlined as follows:

$$L = -\sum_{i=1}^{K} y_i \log(\hat{y_i}) \tag{18}$$

$$Accuracy = \frac{1}{N}\sum_{i=1}^{N} I(y_i = \hat{y_i}) \tag{19}$$

In accordance with the modeling phase of the 1D CNN-BIGRU network, the final outcome yields precise state predictions.

III. EXPERIMENT

This section outlines the methodology used to examine driver data and subsequently presents the findings of the conducted experiments. The following subsections provide a detailed account of the experimental framework, the dataset employed, the design of the scenarios, and the analysis of the results.

A. Dataset

The present study was conducted using the Kaggle online data science platform. To validate our stress recognition methodology, we utilized four representative physiological signal datasets: DEAP [14], CASE [15], EmoWear [16], and Stress Recognition in Automobile Drivers [17].

The DEAP dataset, established in 2010, comprises multimodal physiological data collected from 32 participants who viewed 40 video clips. This dataset includes various physiological signals, such as electroencephalography (EEG), electromyography (EMG), electrocardiography (ECG), and skin conductance response (SCR), thereby offering a rich multidimensional resource for research in affective computing. In contrast, the CASE dataset emphasizes a comprehensive collection of multimodal emotional signals, acquiring physiological data, including EEG, heart rate variability, and skin conductance response, from participants exposed to diverse stimuli within meticulously designed experimental frameworks. This highlights the breadth of research within the field of affective computing. Furthermore, the EmoWear dataset exemplifies the innovative use of wearable technology for emotion recognition, as it captures physiological indicators such as heart rate, skin conductance, and body temperature through smart wearable devices, facilitating continuous monitoring of emotional states in everyday contexts. Lastly, the Stress Recognition in Automobile Drivers dataset is specifically tailored for the identification of stress in drivers, gathering critical physiological signals, including electrocardiograms, skin conductance responses, and respiratory rates, in real-world driving scenarios. This dataset provides direct and precise data support for the analysis of stress related to driving.

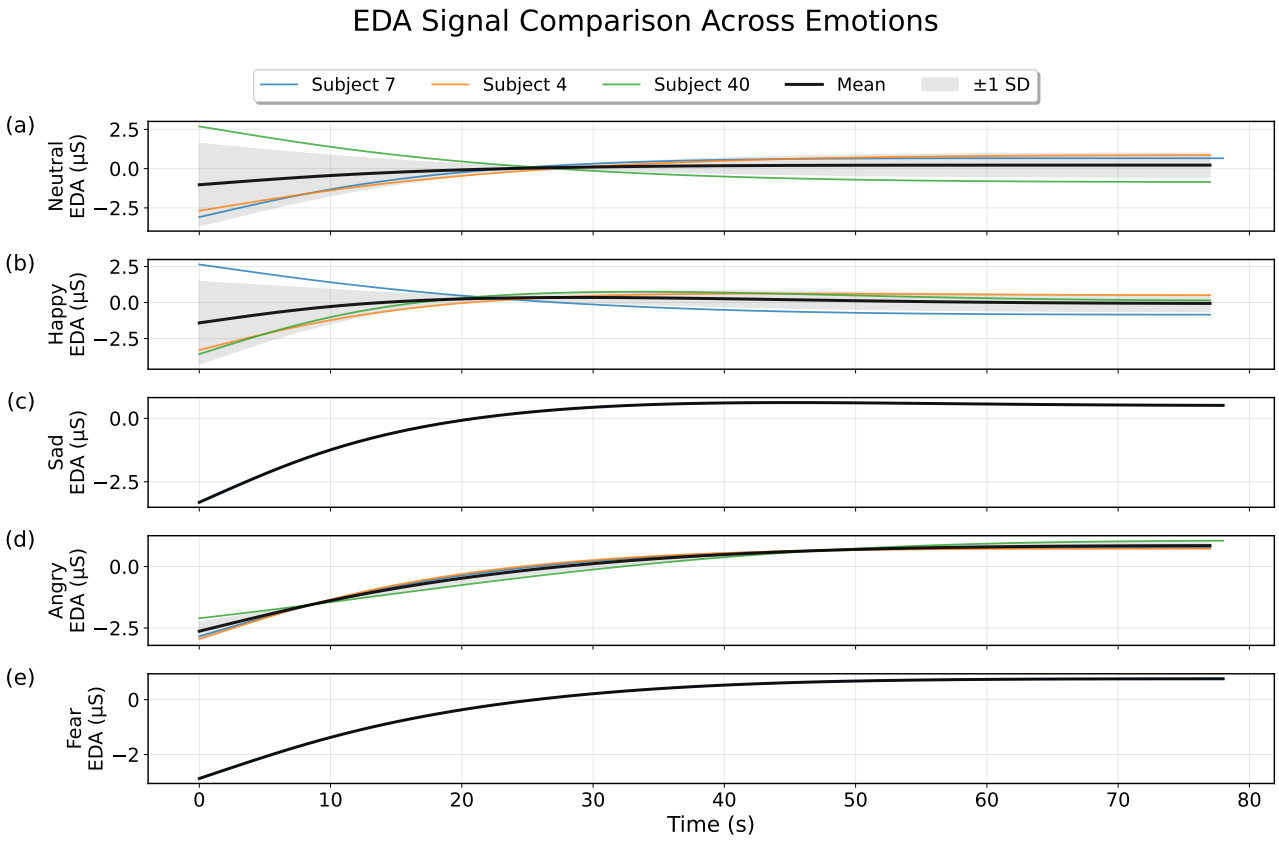

Fig. 4. Emotional Comparative Analysis of Physiological Signals.

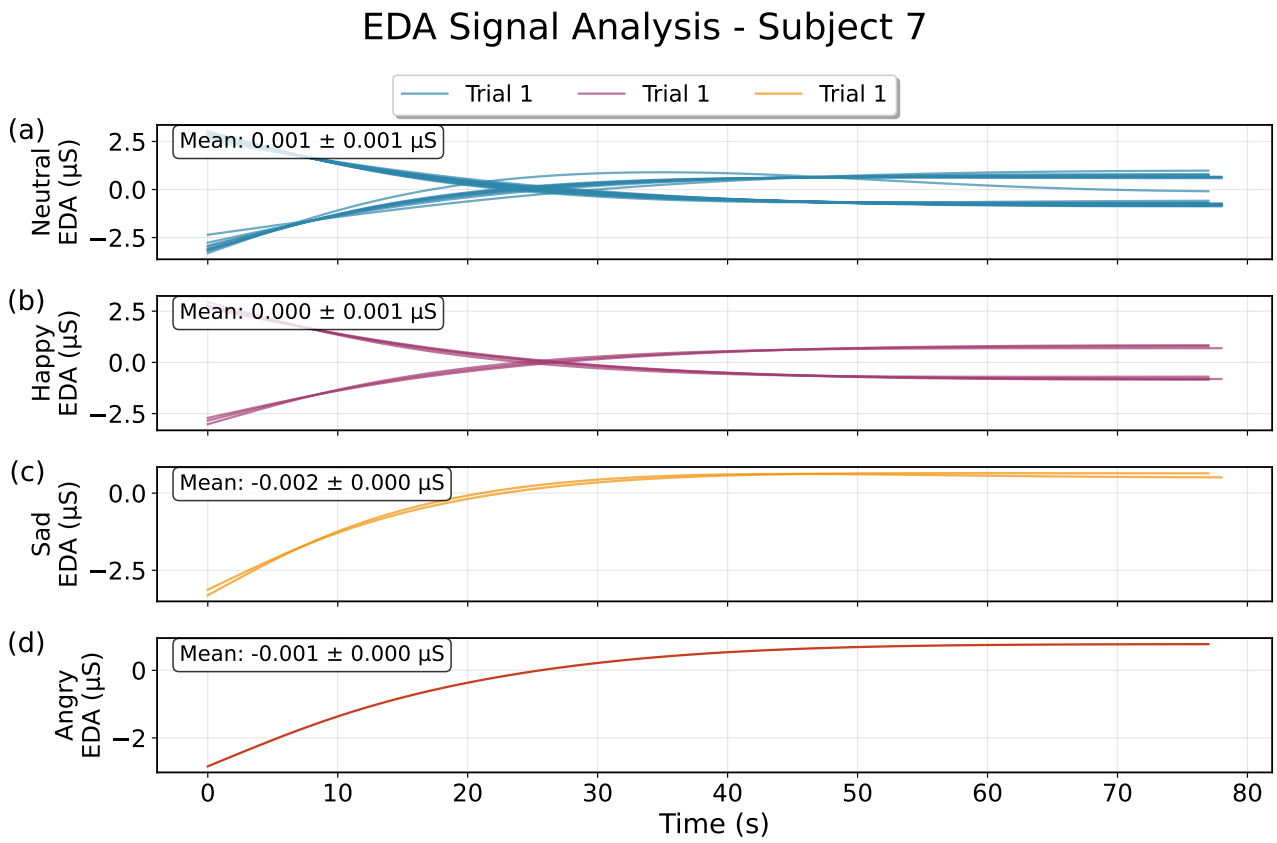

Fig. 5. Single Subject Analysis.

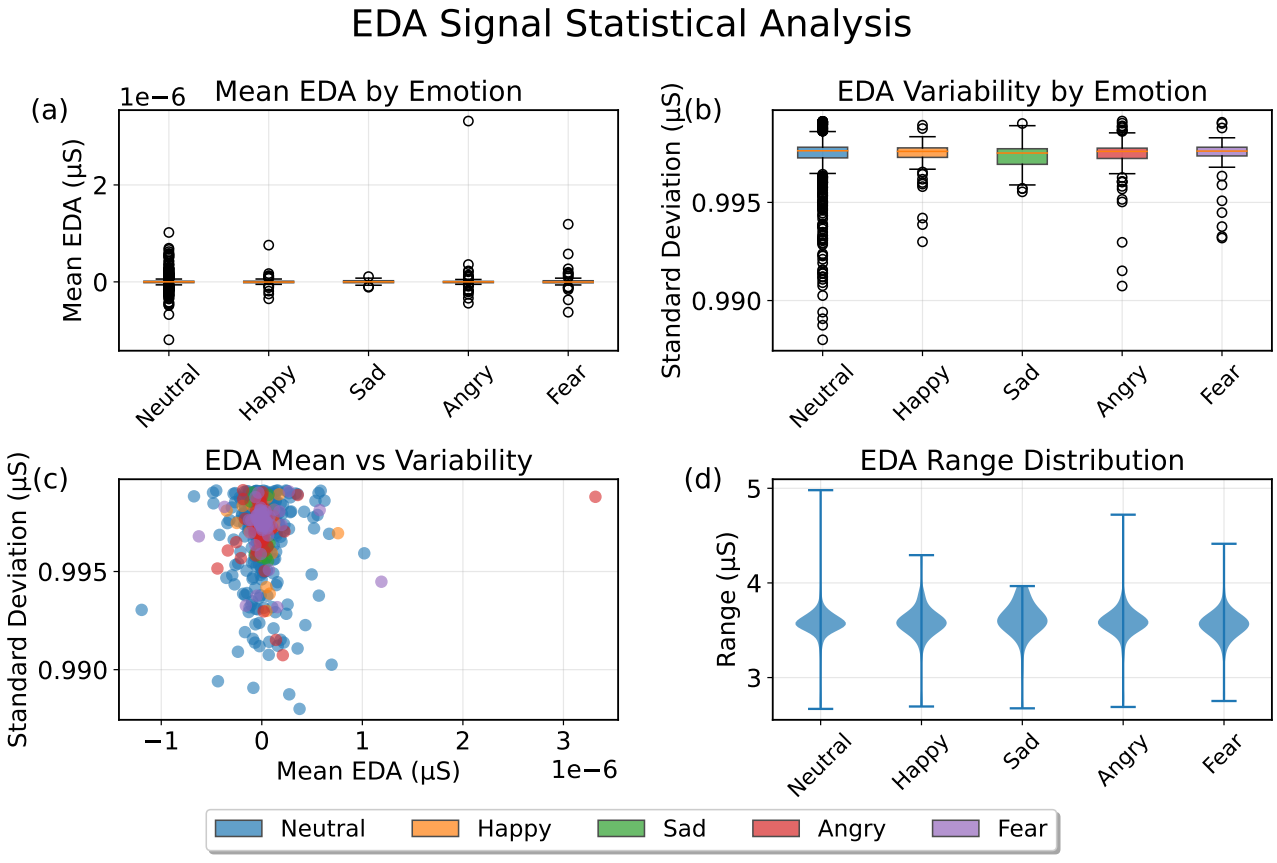

Fig. 6. Statistical Characteristics Analysis.

B. Comparison Experiment

This study conducted a systematic examination of electrodermal activity (EDA) signals to elucidate their unique characteristics across various emotional states, specifically neutral, pleasant, sad, angry, and fearful conditions. As illustrated in Figure 4, EDA signals corresponding to different emotional states reveal significant variations in their temporal characteristics. A comparative analysis of signals from multiple participants indicated that, despite individual physiological baseline differences, the trajectories of signal changes across emotional categories exhibited a high degree of consistency among participants. Statistical evaluations demonstrated that the amplitude of EDA responses during pleasant and angry emotional states was significantly greater than that observed in neutral conditions, while sad emotions were associated with comparatively lower activation levels. The mean curves, along with their standard error of the mean (SEM) and standard deviation (SD) intervals, suggest that EDA signals across different emotional states possess distinct statistical characteristics, thereby establishing a physiological basis for subsequent multimodal feature fusion.

To evaluate the reproducibility and stability of electrodermal activity (EDA) signal measurements, this study conducted a comprehensive analysis of multiple repeated measurements from a single participant exposed to identical emotional stimuli. Figure 5 illustrates the EDA signal change patterns for this participant across neutral (36 trials), pleasant (7 trials), sad (2 trials), and angry emotional conditions. Statistical analysis reveals that the standard error of the mean for signals across all emotional categories remained within 0.001 μS, with a mean of 0.001 ± 0.001 μS for the neutral state, 0.000 ± 0.001 μS for sadness (-0.002 ± 0.000 μS), and anger (-0.001 ± 0.000 μS). The high consistency observed in the multi-trial data highlights the stability of the EDA signal acquisition system and the effectiveness of the employed signal preprocessing methods. Furthermore, it affirms the reproducibility of the emotion-induced paradigm, thereby providing a robust data foundation for subsequent pattern recognition algorithms.

Further statistical feature analysis of electrodermal activity (EDA) signals elucidates the quantitative distinctions among various emotional states. Figure 6a indicates that the mean distribution range for each emotional category spans from -3 to 3 μS, with fear demonstrating the highest mean response and sadness reflecting a negative shift. The standard deviation analysis presented in Figure 6b reveals that the signal variability for each emotional state ranges from 0.988 to 0.998 μS, with fear and anger exhibiting relatively high fluctuations. The mean-variability scatter plot in Figure 6c illustrates the distribution patterns of different emotions within a two-dimensional feature space, where each emotional category forms relatively independent clustering regions. Additionally, the signal range analysis depicted in Figure 6d indicates that the dynamic range for each emotional state is distributed between 3.0 and 5.0 μS, with fear displaying the most extensive signal variation range. These systematic differences in statistical features provide a significant feature space for emotion recognition based on EDA signals, further demonstrating the effectiveness and feasibility of utilizing physiological signals in emotion computing applications.

Figure 7 illustrates the characteristics of time-domain waveforms for four physiological signals—electrodermal activity (EDA), blood volume pulse (BVP), electrocardiogram (ECG), and skin temperature (SKT)—across three distinct emotional states: neutral, happy, and sad. The findings reveal that these physiological signals exhibit varying levels of sensitivity to emotional states. Specifically, EDA signals show an increased baseline during emotional activation, BVP signals display dynamic patterns associated with different emotions, ECG signals reflect emotion-dependent variations in heart rate variability, and SKT signals maintain a relatively stable baseline with emotion-specific alterations. These variations in waveform characteristics establish a significant physiological basis for recognizing emotions through multimodal approaches.

Figure 8 evaluates the efficacy of four physiological signals in emotion classification from a statistical feature analysis perspective. Panel (a) presents box plots that indicate significant differences in the mean distributions of each signal across various emotional states. Panel (b) features histograms that illustrate the variability distributions, revealing that skin temperature (SKT) signals exhibit relatively concentrated variability characteristics. Panel (c) includes a correlation matrix that demonstrates a strong negative correlation between electrodermal activity (EDA) and SKT (r = -0.94), as well as a positive correlation between electrocardiogram (ECG) and SKT (r = 0.71), highlighting the interdependent relationships among different physiological systems. Panel (d) presents a separability analysis, indicating that SKT signals exhibit the highest discriminative capability for emotion classification (separability = 1.75), followed by EDA signals (separability = 1.17). These findings provide quantitative criteria for feature selection in the development of emotion recognition systems that utilize the fusion of multimodal physiological signals.

In preliminary experiments, we employed three feature extraction techniques—graph signal processing, topological decomposition, and empirical mode decomposition (EEMD)—in conjunction with conventional machine learning classifiers, including k-nearest neighbors, decision trees, and naive Bayes, to classify driver states. The results indicate significant discrepancies in classification performance across various datasets when different feature extraction methods are utilized. Data obtained from experiments conducted on the DEAP, CASE, EmoWear, and Stress Recognition in Automobile Drivers datasets, as illustrated in Table 1, highlight the performance metrics of decision trees and k-nearest neighbors. It was noted that the classification accuracy achieved through traditional feature engineering and machine learning methodologies was relatively modest, with the highest accuracy recorded at only 68.5%. This observation suggests that reliance on a single feature extraction technique and conventional classifiers is insufficient for effectively capturing the complex physiological signal patterns exhibited by drivers.

In light of the limitations associated with conventional feature engineering and classification techniques, we investigated the efficacy of neural networks for feature extraction. By employing graph signal processing,

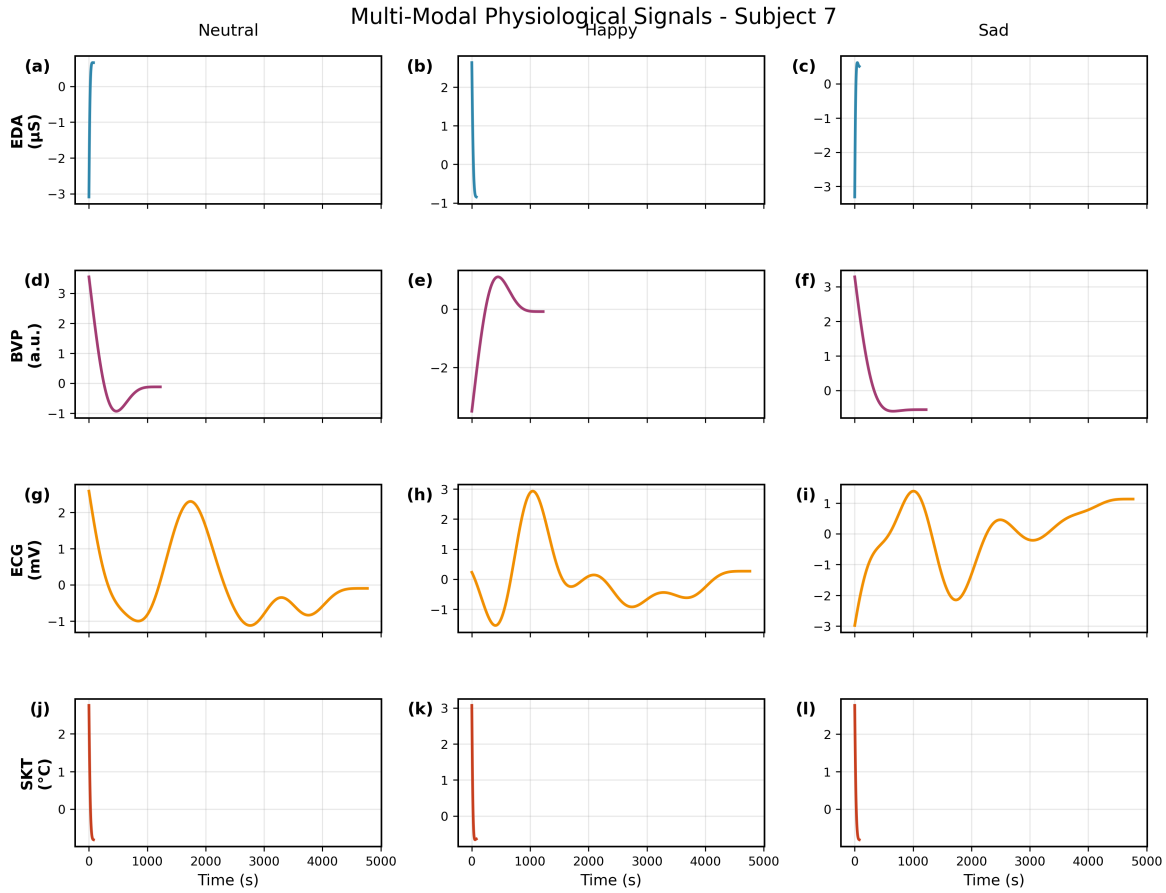

Fig. 7. Multimodal Physiological Signal.

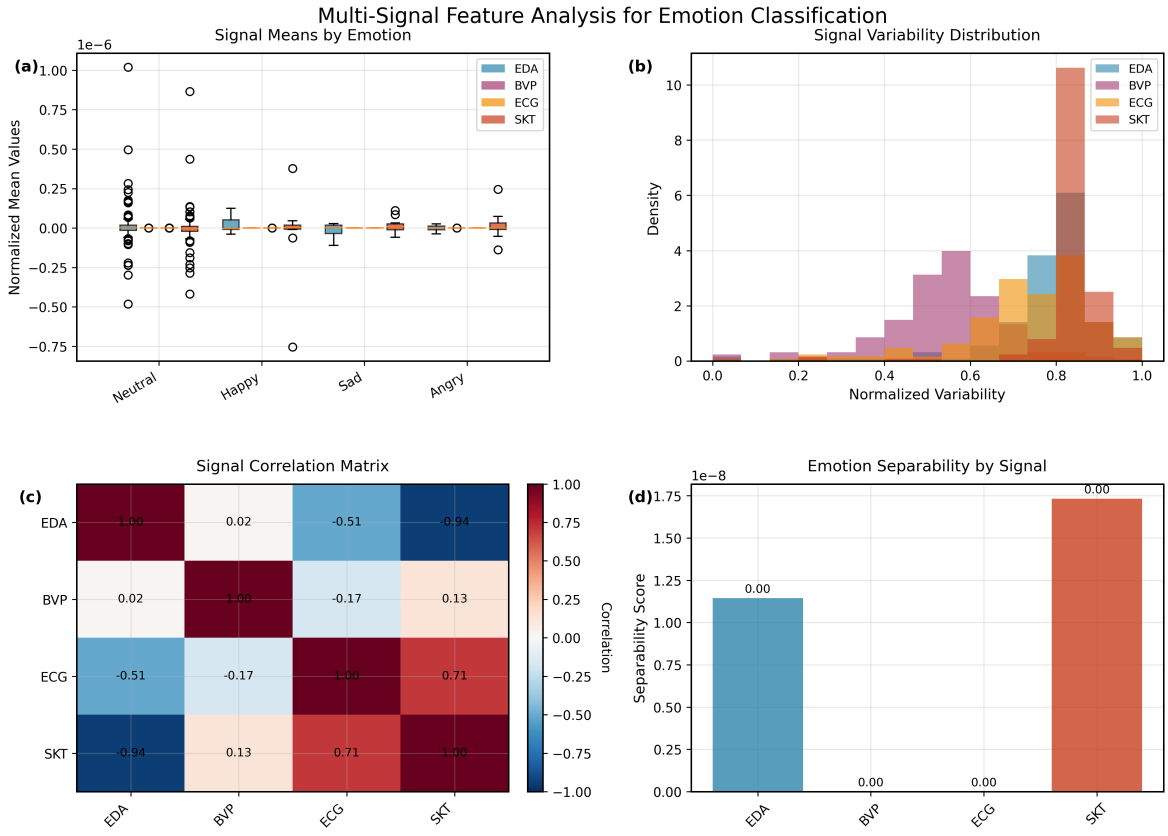

Fig. 8. Multimodal Feature Analysis for Emotion Classification.

TABLE I
TRADITIONAL CLASSIFICATION RESULTS AFTER THREE FEATURE EXTRACTIONS FOR DIFFERENT SIGNALS

Signal	Balanced Accuracy	F1
EDA	0.753 (0.769)	0.663 (0.679)
ECG	0.756 (0.769)	0.676 (0.678)
BVP	0.755 (0.762)	0.672 (0.675)
SKT	0.755 (0.769)	0.677 (0.679)

topological decomposition, and empirical mode decomposition (EEMD) in conjunction with convolutional neural networks (CNNs) and bidirectional gated recurrent units (BiGRUs), we present a comparative analysis of the two neural network architectures in Table 2. The findings indicate a significant enhancement in classification performance. Specifically, the classification accuracy achieved through various feature extraction techniques across different neural network frameworks consistently surpasses that of traditional machine learning approaches, with the highest recorded accuracy reaching 75.3%. This outcome underscores the advantages of neural networks in the domain of automatic feature learning and extraction, while also highlighting the substantial impact of physiological signal feature extraction methods on classification efficacy.

TABLE II
NEURAL NETWORK CLASSIFICATION RESULTS AFTER THREE FEATURE EXTRACTIONS FOR DIFFERENT SIGNALS

Signal	Balanced Accuracy	F1
EDA	0.765 (0.781)	0.658 (0.676)
ECG	0.763 (0.775)	0.656 (0.685)
BVP	0.765 (0.786)	0.662 (0.674)
SKT	0.764 (0.781)	0.673 (0.686)

To enhance the precision of driver state recognition, we propose a novel feature fusion strategy. This approach involves the parallel integration of three feature extraction techniques: graph signal processing, topological decomposition, and empirical mode decomposition (EEMD). These techniques are subsequently combined with a hybrid architecture that includes a one-dimensional convolutional neural network and a bidirectional gated recurrent unit (1D CNN-BiGRU). Our findings indicate that this fused feature extraction method significantly improves the accuracy of driver state recognition, achieving an accuracy rate of 82.7% across four datasets. This represents a substantial enhancement in performance compared to individual feature extraction methods and conventional classification techniques. The synergistic combination of parallel multi-feature extraction and deep learning models offers a promising new methodology for analyzing complex physiological signals.

C. Analysis of Results

This research has made significant advancements in the field of driver state recognition by utilizing a combination of multi-signal feature extraction techniques and deep learning methodologies. The experimental findings, as illustrated in Table 3, highlight the exceptional performance of the proposed 1D CNN-BiGRU network across four publicly available datasets. In comparison to traditional machine learning approaches, such as decision trees, k-nearest neighbors, and naive Bayes, our model demonstrates a substantial improvement in accuracy. Furthermore, the superiority of this method is supported by a comparative analysis with contemporary mainstream deep learning architectures. Specifically, when compared to a standalone CNN architecture, our approach achieves accuracy improvements of 6%, 6%, 6%, and 13% on the CASE, DEAP, PSY, and EMO datasets, respectively. In comparison to the BiGRU network, the accuracy enhancements are 2%, 3%, 4%, and 11%, respectively. Importantly, comparative experiments with the currently prevalent Transformer architecture reveal that, despite the theoretical advantages of Transformers in sequence modeling, the proposed 1D CNN-BiGRU fusion architecture demonstrates superior adaptability and stability in multimodal physiological signal processing tasks, outperforming the Transformer model by an average margin of 2 to 4 percentage points across the four datasets. These comparative results validate the effectiveness of multi-signal feature extraction and innovative deep-learning network architectures, while also emphasizing the importance of architecture optimization tailored to specific application contexts.

In comparison to temporal convolutional networks (TCNs), which have demonstrated remarkable efficacy in modeling time-series data in recent years, the methodology presented in this paper offers significant advantages. Specifically, it achieves a 5% improvement in accuracy on the CASE and DEAP datasets, a 1% increase on the PSY dataset, and a 7% enhancement on the EMO dataset. Although TCNs possess commendable parallelization capabilities and effectively manage receptive fields to address long-term dependencies, their performance is inferior to the proposed 1D CNN-BIGRU architecture, particularly in the context of the complex feature fusion task associated with multimodal physiological signals. These comparative findings validate the effectiveness of multi-signal feature extraction and the innovative deep learning network architectures employed, while also highlighting the importance of optimizing architectural design for specific application contexts.

Figure 9 presents the confusion matrix for driver state detection on the DRIVEDB dataset, which encompasses three categories: normal state (0.0), mild fatigue (1.0), and severe fatigue (2.0). The model accurately predicted 200,145 samples in the normal state category, 755,663 samples in the mild fatigue category, and 341,370 samples in the severe fatigue category. These outcomes provide valuable insights for future model optimization efforts.

The technological advancements presented in this study primarily stem from the synergistic optimization of several critical components. Initially, the representation of physiological signals was enhanced through the application of sophisticated extraction methodologies, including graph signal processing, topological decomposition, and empirical mode decomposition (EEMD). The multidimensional integration of physiological signals—specifically electrodermal activity (EDA), electrocardiogram (ECG),

TABLE III
RESULTS OF SPATIO-TEMPORAL FEATURE EXTRACTION IN FIVE NEURAL NETWORKS

Model	CASE	DEAP	DRIVEDB	EMOWEAR
CNN	0.92 (0.90)	0.92 (0.94)	0.92 (0.92)	0.84 (0.79)
BIGRU	0.96 (0.94)	0.95 (0.91)	0.94 (0.90)	0.86 (0.83)
Transformer	0.95 (0.96)	0.94 (0.94)	0.97 (0.96)	0.92 (0.93)
TCN	0.93 (0.94)	0.91 (0.95)	0.95 (0.96)	0.89 (0.9)
1D CNN-BIGRU	0.98 (0.95)	0.98 (0.96)	0.98 (0.96)	0.97 (0.94)

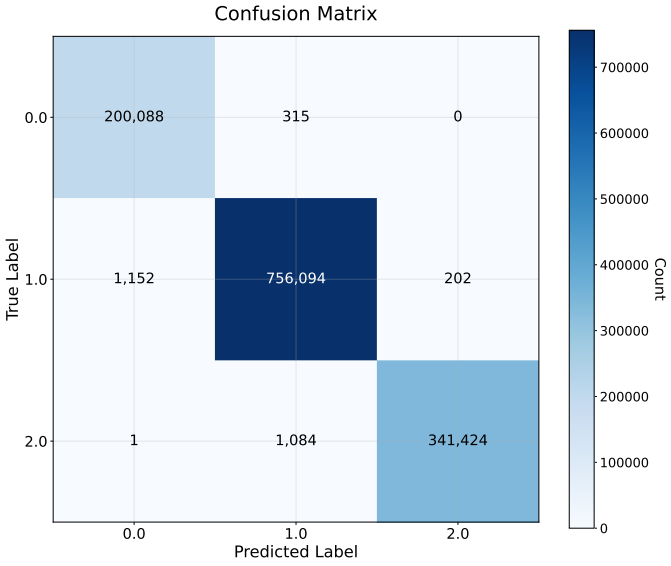

Fig. 9. Confusion Matrix.

blood volume pulse (BVP), and skin temperature (SKT)—produces a more comprehensive and discriminative set of input features for the model. Furthermore, the architecture of the 1D CNN-BiGRU network effectively combines the local feature extraction strengths of convolutional neural networks with the long-range dependency modeling capabilities of bi-directional gated recurrent units. This integration enables the model to more accurately discern the intricate patterns present in driver physiological signals. Additionally, the joint training conducted on four distinct datasets significantly enhances the model's generalization capacity and robustness, thereby providing substantial technical support for the recognition of driver states across various scenarios.

IV. CONCLUSION

This study presents a novel method for detecting driver behavior states that integrates spatiotemporal feature extraction with a one-dimensional convolutional neural network and bidirectional gated recurrent unit (1D CNN-BiGRU) model, with the goal of enhancing driving safety. The approach employs three distinct feature extraction techniques to effectively capture complex data relationships, thereby improving the representation of both local and global signal characteristics and facilitating the identification of changes in driver states. To classify a range of driver behaviors and emotional responses, a neural network architecture that combines 1D CNN with BiGRU has been developed. This network utilizes various features and structural components, including deep convolution, feature fusion, and attention mechanisms, to achieve an optimal balance between complexity and accuracy. The robustness of the proposed method is strengthened by the use of multiple datasets, which encompass both dynamic and static emotional states, as well as diverse data types for activity classification. Simulation results demonstrate that the proposed deep learning model outperforms existing alternatives in terms of reducing complexity while enhancing accuracy.

REFERENCES

[1] H. Zhou and H. Dai, "Research on fatigue driving detection based on deep learning." *Engineering Letters*, vol. 33, no. 2, pp. 348–356, 2025.
[2] C. TAOUSSI, I. HAFIDI, and A. METRANE, "Machine learning and deep learning in health informatics: Advancements, applications, and challenges." *Engineering Letters*, vol. 33, no. 5, pp. 1448–1461, 2025.
[3] G. Xiang, S. Yao, H. Deng, X. Wu, X. Wang, Q. Xu, T. Yu, K. Wang, and Y. Peng, "A multi-modal driver emotion dataset and study: Including facial expressions and synchronized physiological signals," *Engineering Applications of Artificial Intelligence*, vol. 130, p. 107772, 2024.
[4] P. Lalwani and G. Ramasamy, "Human activity recognition using a multi-branched cnn-bilstm-bigru model," *Applied Soft Computing*, vol. 154, p. 111344, 2024.
[5] D. T. N. El-Deen, R. S. El-Sayed, A. M. Hussein, and M. S. Zaki, "Multi-label classification for sentiment analysis using cbga hybrid deep learning model." *Engineering Letters*, vol. 32, no. 2, pp. 340–349, 2024.
[6] Y. Ma, R. Du, A. Abdelraouf, K. Han, R. Gupta, and Z. Wang, "Driver digital twin for online recognition of distracted driving behaviors," *IEEE Transactions on Intelligent Vehicles*, vol. 9, no. 2, pp. 3168–3180, 2024.
[7] Q. Wei and K. Yang, "Research on interpretable recommendation algorithms based on deep learning." *Engineering Letters*, vol. 32, no. 3, pp. 560–568, 2024.
[8] J. Chen, X. Lin, W. Ma, Y. Wang, and W. Tang, "Eeg-based emotion recognition for road accidents in a simulated driving environment," *Biomedical Signal Processing and Control*, vol. 87, p. 105411, 2024.
[9] N. Ying, Y. Jiang, C. Guo, D. Zhou, and J. Zhao, "A multimodal driver emotion recognition algorithm based on the audio and video signals in internet of vehicles platform," *IEEE Internet of Things Journal*, vol. 11, no. 22, pp. 35 812–35 824, 2024.
[10] Z. Hu, H. Wu, and L. He, "A neural network for eeg emotion recognition that combines cnn and transformer for multi-scale spatial-temporal feature extraction." *IAENG International Journal of Computer Science*, vol. 51, no. 8, pp. 1094–1104, 2024.
[11] K. Hong, "Facial expression recognition based on anomaly detection and multispectral imaging." *IAENG International Journal of Computer Science*, vol. 51, no. 10, pp. 1627–1641, 2024.
[12] S. Pain, M. Sarma, and D. Samanta, "Graph signal processing and graph learning approaches to schizophrenia pattern identification in brain electroencephalogram," *Biomedical Signal Processing And Control*, vol. 100, p. 106954, 2025.
[13] Y. R. Veeranki, L. R. M. Diaz, R. Swaminathan, and H. F. Posada-Quintero, "Nonlinear signal processing methods for automatic emotion recognition using electrodermal activity," *IEEE Sensors Journal*, vol. 24, no. 6, pp. 8079–8093, 2024.
[14] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, "Deap: A database for emotion analysis; using physiological signals," *IEEE transactions on affective computing*, vol. 3, no. 1, pp. 18–31, 2011.

[15] K. Sharma, C. Castellini, E. L. Van Den Broek, A. Albu-Schaeffer, and F. Schwenker, "A dataset of continuous affect annotations and physiological signals for emotion analysis," *Scientific data*, vol. 6, no. 1, p. 196, 2019.
[16] M. H. Rahmani, M. Symons, O. Sobhani, R. Berkvens, and M. Weyn, "Emowear: Wearable physiological and motion dataset for emotion recognition and context awareness," *Scientific Data*, vol. 11, no. 1, p. 648, 2024.
[17] J. A. Healey and R. W. Picard, "Detecting stress during real-world driving tasks using physiological sensors," *IEEE Transactions on intelligent transportation systems*, vol. 6, no. 2, pp. 156–166, 2005.