Ingénierie des Systèmes d'Information
Vol. 31, No. 5, May, 2026, pp. 1485-1494
Journal homepage: http://iieta.org/journals/isi

A Hybrid Deep Feature Fusion and Stacked Classifier Network for Monitoring Driver Drowsiness and Distraction

Venkateswarlu Madduri, Venkata Rami Reddy Chirra*

School of Computer Science and Engineering, VIT-AP University, Amaravati 522241, India

Corresponding Author Email: venkataramireddy.chirra@vitap.ac.in

https://doi.org/10.18280/isi.310508

Received: 25 November 2025
Revised: 1 March 2026
Accepted: 15 April 2026
Available online: 31 May 2026

Keywords:
Hybrid Deep Feature Fusion and Stacked Classifier Network, driver monitoring, ensemble learning, feature fusion, deep learning, drowsiness detection, distracted driving, machine learning

ABSTRACT

Drowsiness and distraction remain major contributors to road accidents worldwide. This study presents Hybrid Deep Feature Fusion and Stacked Classifier Network (HDFSNet), a hybrid deep learning framework that integrates multi-Convolutional Neural Network (CNN) feature fusion with stacked ensemble classification to robustly detect subtle driver behaviors, including marginal eye closure, yawning, and small headpose variations. The proposed architecture combines complementary representations from VGG16, ResNet50, and InceptionV3, and employs heterogeneous base classifiers (Support Vector Machine (SVM), Stochastic Gradient Descent (SGD), and Multilayer Perceptron (MLP)) with a logistic regression meta-classifier to improve generalization and reduce inter-class confusion. The model is evaluated on a balanced, post-augmentation merged dataset of 41,094 facial images across eight driver behavior classes, encompassing head pose, eye state, and yawning conditions, using stratified training, validation, and testing splits. Experimental evaluation revealed an overall accuracy of 99.08%, along with macro-averaged precision, recall, and F1 scores exceeding 0.99, and Area Under the Curve (AUC) values of 1.00 for all classes. These results affirm the strength of the introduced feature fusion and ensemble mechanism, demonstrating its potential for accurate, near real-time deployment in driver drowsiness and distraction monitoring applications.

1. INTRODUCTION

Among the most significant contributing factors to road traffic accidents is driver drowsiness and distraction, both of which substantially impair driving performance and heighten the risk of collision, thereby threatening the safety of not only the fatigued or inattentive driver but also other vulnerable road users, including co-passengers and pedestrians. Such incidents can lead to severe consequences, placing countless lives at significant risk. While official statistics often highlight fatality rates, the overall impact of drowsy driving is likely underestimated and may extend well beyond the reported figures. "Drowsy driving," also known as fatigued or tired driving, occurs when a driver operates a vehicle while exhausted or sleepy. As people rely on cars for daily transportation, the risk of serious or fatal collisions with other vehicles or stationary objects significantly increases [1]. Venkateswarlu and Ch [2] highlighted that this condition greatly endangers drivers by raising the chances of drowsiness and the potential for accidents. Distracted driving refers to any instance in which a driver's attentional focus is diverted from the primary driving task toward secondary activities, consequently impairing their judgment, responsiveness, and vehicular control. It is recognized as a significant factor in fatal accidents and injuries on the road [3]. Distracted driving encompasses any secondary activity that interferes with the driver's primary task of safe vehicle operation, mobile device communication, food and beverage consumption, passenger engagement, adjusting in-vehicle controls and navigation systems, and abnormal head-pose orientations. Since many forms of distracted driving can be observed using cameras installed in vehicles, analyzing video frames captured during real-world driving scenarios is a logical approach for detecting such behaviors [4]. Road accidents lead to damage to property and, in some cases, fatalities. There has been a significant rise in accidents caused by distracted driving, which is a growing concern. Distracted driving is generally classified into three categories: visual (looking for items inside the car, checking a GPS, or reading a text message), manual (e.g., eating or drinking while taking hands off the wheel, adjusting the radio, or texting), and cognitive (e.g., losing focus or envisioning, having a conversation with a passenger, or thinking about personal issues). These distractions can result in significant lane deviations, reduced vehicular control, slower reactions to hazards, and a diminished awareness of the roadway surroundings compared to attentive driving [5]. Data published by the National Highway Traffic Safety Administration (NHTSA) reveal that distracted driving contributed to 3,275 fatalities in 2023, highlighting the serious risks posed by inattentive behavior behind the wheel [6]. Drowsy driving is a factor in almost 6,000 fatal accidents each year [7]. A 2024 study by the Insurance Institute for Highway Safety (IIHS) found that users of partial automation were more prone to distractions, such as phone use and eating, than

manual drivers. In addition, drowsy-driving-related crashes caused 633 fatalities in the same year. These statistics underscore the pressing need for heightened public awareness and preventive measures to combat risky driving behaviors. Drowsy driving is a form of impaired driving, contributing to approximately one in five fatal motor vehicle crashes on U.S. roads [8]. Detection methodologies are broadly categorized into 3 primary modalities, encompassing behavioral observations, physiological measurements, and vehicle-derived operational data, which are collectively employed to identify distracted or drowsy driving patterns [9]. Venkateswarlu and Chirra [10] introduced a hybrid Convolutional Neural Network (CNN)-ViT architecture that synergistically combines locally extracted facial features and globally oriented attention mechanisms, achieving highly accurate classification of eye and mouth states for real-time drowsiness detection.

Despite significant progress made in deep learning-based driver monitoring systems, several critical research gaps remain unresolved:

- Limited feature representation: Single CNN models often fail to capture subtle and complex driver behaviors like small head pose changes or micro-expressions.
- Lack of effective feature fusion: Most approaches do not combine complementary spatial, semantic, and contextual features from multiple models, limiting detection accuracy.
- Poor generalization: Models trained on small or imbalanced datasets struggle to handle real-world differences in lighting, facial expressions, and head poses due to limited data augmentation.

To overcome these gaps, we propose Hybrid Deep Feature Fusion and Stacked Classifier Network (HDFSNet), a hybrid deep feature fusion and stacked classifier network that integrates multiple CNNs with ensemble learning, enhancing robustness, accuracy, and generalization for driver drowsiness and distraction detection.

Subtle driver behaviors such as micro-expressions, marginal eye closure, and small head pose deviations are represented at different spatial and semantic levels within facial imagery. Lower-level convolutional layers predominantly capture fine-grained texture and edge-based cues, including eyelid contours, lip boundaries, and localized facial muscle movements, while deeper layers encode higher-level semantic and geometric representations such as global facial structure, head orientation, and contextual pose information. Consequently, single-backbone CNN architectures tend to provide only partial feature representations, which leads to increased inter-class confusion between visually similar driver states, particularly for classes such as MiddleCenter versus MiddleLeft and yawn versus no_yawn.

To empirically validate this limitation, a comparative analysis was conducted between individual CNN-based pipelines and the proposed HDFSNet framework. Models based on single feature extractors (VGG16, ResNet50, and InceptionV3 combined with Support Vector Machine (SVM), Stochastic Gradient Descent (SGD), and Multilayer Perceptron (MLP) classifiers) achieved classification accuracies in the range of 96.89%–98.25%. In contrast, the proposed multi-CNN feature fusion and stacked classification architecture improved overall accuracy to 99.08%, while significantly reducing misclassification rates among subtle head pose and yawning classes.

Furthermore, the classifier stacking strategy functions as a form of decision-level regularization. The heterogeneous base learner model complementary and partially independent decision boundaries within the high-dimensional fused feature space, and the logistic regression meta-classifier learns an optimal combination of their probability outputs. This integration enhances robustness and generalization under variations in illumination, partial occlusion, and driver posture, thereby directly addressing the limitations of limited feature representation, ineffective fusion, and poor generalization identified in the problem definition.

HDFSNet is novel because it combines features from three different CNN models (VGG16, ResNet50, and InceptionV3) and uses multiple classifiers (SVM, SGD, MLP) stacked together with a meta-classifier for final decisions. This approach captures more varied information and improves accuracy compared to methods using a single model or classifier. It also uses automated facial region detection and data augmentation to enhance generalization. Overall, HDFSNet offers a robust, accurate, and real-time solution for detecting driver drowsiness and distraction.

- Introducing an innovative architecture that combines CNN-driven hierarchical feature extraction with ML classifiers to identify driver drowsiness and distraction.
- Data augmentation strategies were employed to mitigate class imbalance, thereby improving model robustness and generalization.
- Two datasets, encompassing four classes each for drowsiness and distraction, were integrated to enhance the classifier's performance and ability to generalize.
- For each feature extraction architecture — VGG16, ResNet50, and InceptionV3—separate classifiers, including SVM, SGD, and MLP, were individually applied to perform the classification task.
- A hybrid model, HDFSNet, was developed by integrating VGG16, ResNet50, and InceptionV3 for feature extraction with stacked SVM, SGD, and MLP classifiers, and a meta-classifier, logistic regression, is used for final prediction.
- To evaluate the proposed method, standard classification metrics encompassing precision, recall, F1-score, accuracy, confusion matrix, Receiver Operating Characteristic (ROC), and Precision-Recall (PR) curves were employed.

The literature on drowsiness and distraction detection is reviewed in Section 2. Section 3 introduces the proposed approach and provides a detailed description of the system's components. In Section 4, the performance of each CNN architecture is evaluated and the experimental results are discussed. Finally, Section 5 concludes the manuscript by summarizing the key findings and outlining directions for future research.

2. RELATED WORK

Hssayeni et al. [11] explored two approaches for distracted driver detection using dashboard camera images: (1) traditional handcrafted features (HOG, SIFT) combined with SVM classifiers, and (2) deep learning with transfer learning CNNs like AlexNet, VGG-16, and ResNet-152. While

handcrafted features with SVMs achieved low accuracy (27.7%), in-tuned CNNs performed significantly better, with ResNet-152 reaching 85% and VGG-16 reaching around 82.5%. Replacing CNN classifiers with SVMs did not improve results, highlighting the superior feature learning capabilities of deep CNNs.

Omerustaoglu et al. [12] designed a two-phase multimodal deep learning system for distracted driver detection by combining visual and sensor data. They collected a new dataset with driver images and sensor inputs (gyroscope, accelerometer, GPS, OBD) and used a public distracted driver dataset for transfer learning. The first stage uses CNNs to analyze images, while the second stage employs Long Short-Term Memories (LSTMs) to model sequential sensor data. Fusion methods at feature and prediction levels integrate CNN and LSTM outputs, improving accuracy to 85%, up from 76% with vision-only models, while reducing false positives and enhancing robustness.

Reddy et al. [13] introduced a system designed to automatically detect driver drowsiness. This approach involved image preprocessing, face and eye detection using the Viola-Jones algorithm, and feature extraction through the Discrete Wavelet Transform (DWT). A Radial Basis Function Neural Network (RBFNN) was then used to classify the extracted features and attained 95.4% accuracy in detecting alert versus drowsy conditions.

Chillakuru et al. [14] developed a three-stage distracted driver detection model combining deep learning and traditional methods. It uses CNNs (VGG16, Inception, Xception) to extract deep features along with texture and motion features, refined via PCA. Classification is done with a Deep Belief Network (DBN) and Radial Basis Function (RBF) network, optimized by Adaptive Gradient-based Optimizer (AGBO). A final ranking assesses distraction levels. The model achieves 97.67% accuracy, 97.68% sensitivity, and 97.55% precision, outperforming traditional methods by up to 67% in K-fold validation, though the F1-score could still improve.

Hasan et al. [15] employed K-fold and stratified K-fold cross-validation, along with Leave-One-Participant-Out (LOPO) validation, to evaluate model performance while addressing data imbalance and individual variability. Data from 26 participants were used, with 22 physiological features (EEG, EOG, ECG) extracted over 5-second epochs. From 9360 observations, 156 were selected for analysis (79 drowsy, 77 awake). Using explainable machine learning classifiers, particularly random forest, the system achieved 80.1% accuracy, 82.2% specificity, and 70.3% sensitivity in detecting drowsiness.

Lamouchi et al. [16] addressed driver drowsiness detection using supervised machine learning on the UTA-RLDD dataset, which contains real drowsy driving videos. They extracted spatiotemporal facial features using LBP-TOP and reduced feature dimensionality before classifying with a linear SVM. For multi-class sleepiness detection, a one-versus-all strategy was used. The model achieved 82% accuracy for binary classification on UTA-RLDD, but accuracy dropped to 61% when a "Low-Vigilance" class was added, highlighting the challenge of distinguishing subtle drowsiness levels. On the DROZY dataset, however, performance was higher, with 90% accuracy, indicating the method's robustness for use in Advanced Driver Assistance Systems (ADAS).

3. METHODOLOGY

3.1 Transfer learning based model for feature extraction

The system combines three pre-trained CNNs (VGG16, ResNet50, and InceptionV3) with three classifiers (SVM, SGD, MLP) to leverage their complementary strengths in feature extraction and decision-making. Each CNN extracts feature vectors from driver images, which are then classified using the chosen algorithm, resulting in nine hybrid configurations: VGG16-SVM, VGG16-SGD, VGG16-MLP, ResNet50-SVM, ResNet50-SGD, ResNet50-MLP, InceptionV3-SVM, InceptionV3-SGD, and InceptionV3-MLP.

3.1.1 VGG16 for feature extraction

The VGG16 [17] model is a simple and effective CNN known for strong image classification performance. It uses stacked 3 × 3 convolutional layers and 2 × 2 max-pooling to capture fine image details. In this work, VGG16 is used without its top layers and applies Global Average Pooling (GAP) to produce a compact feature vector, which is then classified by SVM, SGD, and MLP. This approach provides robust features for accurately detecting driver states from facial cues.

3.1.2 ResNet50 for feature extraction

ResNet50 [18], comprising 50 convolutional layers, introduced the residual connections to improve training and capture complex features. In this system, ResNet50's top layer is removed and GAP is used to create a fixed-size feature vector, which helps accurately distinguish subtle driver behaviors. These features are then classified by base models, enhancing overall robustness and generalization.

3.1.3 InceptionV3 for feature extraction

InceptionV3 [19] is a CNN designed for high accuracy and efficiency, using parallel filters of different sizes for multiscale feature extraction. Here, the model's top layers are removed, and GAP is applied to produce a feature vector. This approach helps identify complex driver behaviors, with extracted features used for final classification.

3.2 Base classifiers

3.2.1 Support Vector Machine

SVM is a supervised classification technique that partitions data into classes by constructing a suitable hyperplane within a high-dimensional feature space. In this work, deep features from CNNs are input to the SVM, which uses a polynomial kernel for nonlinear separation, improving class distinction. The decision function of SVM can be defined as shown in Eq. (1).

$$f(x_j) = \sum_{m=1}^{N} \alpha_m y_m K(x_m, x_j) + b \quad (1)$$

Here, the input features are evaluated against the support vectors x_m—which represent the critical deep features—using the kernel function K. The final weighted result includes the influence of the Lagrange multipliers α_m, the training labels y_m, and the bias b.

The polynomial kernel function utilized in this implementation is as follows in Eq. (2).

$$K(x_m, x_j) = (\gamma . x_m^{\tau} x_j + r)^d \quad (2)$$

where, γ is the scaling factor, r is the independent term, and d represents the polynomial degree of the kernel.

3.2.2 Stochastic Gradient Descent

SGD is an efficient streamlining method that updates model parameters using single or small batches of data, reducing memory use and speeding up training. Here, an SGD classifier with logistic regression loss was applied to deep features from VGG16, ResNet50, and InceptionV3 to quickly and effectively classify driver states. The parameter modification rule for SGD is defined in Eq. (3).

$$\theta_{s+1} = \theta_s - \eta . \Delta_\theta L(\theta_s; x_j, y_j) \quad (3)$$

The update rule for the model parameters is defined as follows: θ_s denotes model parameters at iteration s, η, learning rate, which governs the magnitude of parameter adjustments during optimization, $L(\theta_s; x_j, y_j)$, loss function evaluated on j-th training sample (x_j, y_j), $\nabla_\theta L$ denotes the gradient of the loss about the parameters.

3.2.3 Multilayer Perceptron

The MLP, a foundational feed-forward neural network architecture, is characterized by one or more intermediate layers positioned between the input features and output predictions. The potential to model complex associations within the data is introduced through the use of non-linear activation functions, and each layer is entirely connected to the next. The MLP processes input features through multiple layers of transformation, as shown mathematically for two hidden layers in Eq. (4).

$$\hat{y} = f(W_3 . \sigma_2(W_2 . \sigma_1(W_1 . x + b_1) + b_2) + b_3) \quad (4)$$

where,

- x is an input feature vector (here, deep features extracted from CNNs).
- W_1 denotes weight matrices for the first hidden layer, W_2 is for the second hidden layer, and W_3 for the output layer, respectively.
- b_1, b_2, b_3 represent bias vectors for corresponding layers.
- σ_1, σ_2 are activation functions like ReLU or tanh.
- $f(\cdot)$ represents the final activation function for output (e.g., softmax for classification).
- $\hat{y}$ is the predicted output vector.

3.3 Hyperparameter configuration for base classifiers

3.3.1 Support Vector Machine

A polynomial kernel function defined in Eq. (2) where the scaling factor is set to γ = 0.01, the independent term is r = 1, and the polynomial degree is d = 3. The regularization parameter is fixed at C = 1.0.

3.3.2 Stochastic Gradient Descent

The SGD classifier is trained using a logistic regression loss function with an altered learning rate of η = 0.001. L2 regularization is applied with α = 0.0001, and the iteration limit is set to a maximum of 1000. Training is performed using mini-batches of size 64 to balance convergence stability and computational efficiency.

3.3.3 Multilayer Perceptron

The MLP architecture consists of two fully connected hidden layers with 512 and 128 neurons, respectively, using ReLU activation functions for non-linear feature transformation. The output layer employs a softmax activation function for probabilistic multi-class classification. The network is optimized using the Adam optimizer with a learning rate of 0.001, and training is performed for a maximum of 200 epochs. To prevent overfitting, early stopping with a patience of 10 epochs based on validation loss is applied.

3.4 Proposed Hybrid Deep Feature Fusion and Stacked Classifier Network framework

Figure 1, depicted in the diagram, outlines a comprehensive multi-stage ensemble learning framework designed to effectively detect driver distraction and drowsiness. This system combines deep learning-based feature extraction with traditional machine learning classifiers, enhancing both accuracy and robustness. It consists of five key modules: preprocessing, feature extraction, feature fusion, base classification, and meta-classification. Each module serves a crucial role in improving the system's generalization and predictive performance. By analyzing driver images, the framework accurately identifies various states related to driver attention and fatigue—such as head position, eye state, and yawning—through a structured stacking approach that leverages the benefits of deep learning with those of machine learning methods.

3.4.1 Input and preprocessing module

The system accepts input images of driver faces, rescaled to a standard spatial resolution of 224 × 224 pixels to ensure seamless compatibility with transfer learning CNNs. To enhance dataset diversity and mitigate class imbalance, real-time data augmentation is applied to generate balanced samples for each class. These enhancements include random transformations, namely rotational perturbations (±15°), horizontal image flipping, and brightness scaling (0.7–1.3), which simulate real-world variations in lighting, head tilts, and gaze directions.

Such data augmentation approaches play an essential and crucial role in maintaining the robustness of the model under varying lighting conditions, occlusions, and driver postures by exposing it to a wide spectrum of visual scenarios during training. In addition to this, the system employs automated facial landmark identification via Dlib's 68-point algorithm, which ensures consistent and precise localization of key facial regions, eyes, mouth, and head pose, even when the driver's posture varies or partial occlusion occurs.

This preprocessing pipeline, combined with the subsequent fusion of various feature representations from three complementary CNN architectures and a stacked ensemble of classifiers, significantly enhances the system's resilience and generalization capabilities. Consequently, the model reliably detects driver drowsiness and distraction under diverse and challenging real-world driving conditions.

- VGG16: Uses stacked 3 × 3 convolutions to capture fine textures, ideal for detecting subtle eye and facial changes indicating drowsiness or distraction.
- ResNet50: Utilizes residual connections to extract deeper semantic features, helping distinguish visually similar driver states like head orientations.

- InceptionV3: Applies multi-branch convolutions for multi-scale analysis, capturing complex facial poses and cognitive load indicators.

Each model generates a high-dimensional feature vector offering a unique perspective, and these complementary features are fused for robust and accurate driver behavior classification.

3.4.2 Feature fusion module

After extracting feature vectors independently from VGG16, ResNet50, and InceptionV3, the Feature Fusion Module integrates these diverse representations into a unified feature vector for each image. Specifically, the outputs from each CNN—corresponding to the training and validation sets—are concatenated along the feature dimension using numpy.concatenate() function. This horizontal stacking guarantees the preservation and collective utilization of each model's complementary qualities.

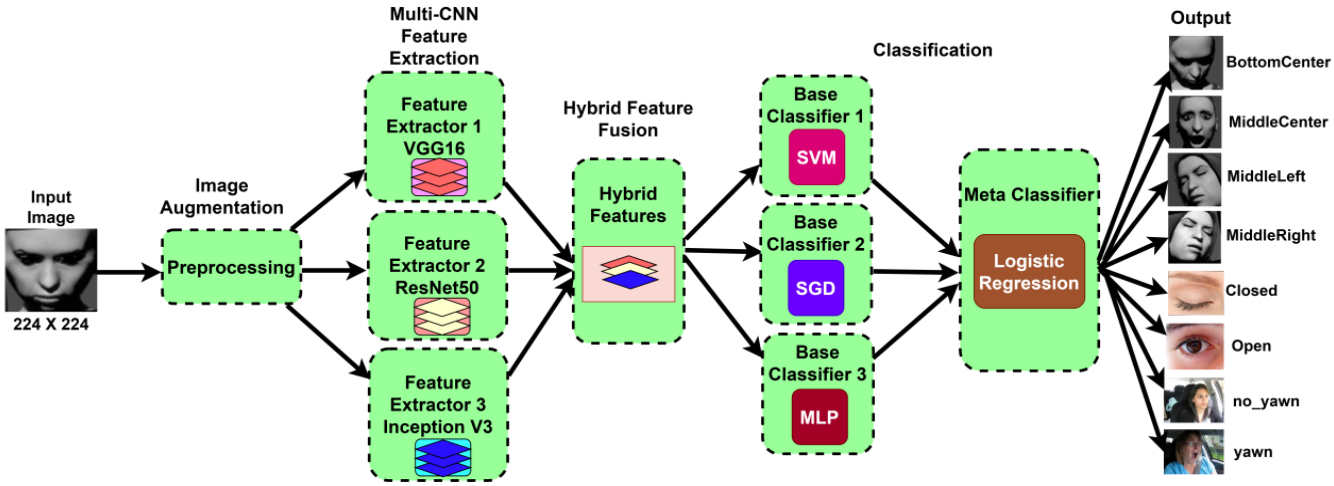

Figure 1. Simplified data flow of the proposed Hybrid Deep Feature Fusion and Stacked Classifier Network (HDFSNet) architecture

- VGG16 contributes low-level texture and edge information.
- ResNet50 provides mid- to high-level semantic features due to its deep residual structure.
- InceptionV3 offers multi-scale contextual insights through its parallel convolutional filters.

The resulting fused feature vector captures a rich blend of spatial, contextual, and semantic attributes, significantly enhancing the system's capacity to differentiate between various driver behaviors. This multi-perspective representation is critical for robust and generalized classification in real-world scenarios.

The deep feature vectors extracted independently from VGG16, ResNet50, and InceptionV3 are subjected to L2 normalization prior to fusion to ensure consistent scaling across heterogeneous feature spaces and to prevent any single backbone from dominating the fused representation due to differences in feature magnitude. In the primary experimental configuration, no dimensionality reduction is applied, as preserving the full high-dimensional feature space is critical for retaining fine-grained discriminative cues associated with subtle driver behaviors such as marginal eye closure, yawning, and head pose deviations. After normalization, the feature vectors are concatenated along the feature dimension to form a unified fused feature representation for subsequent classification.

3.4.3 Base classifiers module

The fused feature vector obtained from the feature fusion stage is fed into an ensemble of three distinct base classifiers, each designed to depict different patterns in the data and contribute uniquely to the overall decision process:

- SVM: Utilizes a polynomial kernel to map input features to higher dimensions, allowing it to find the most effective hyperplanes that distinguish classes with optimal margin—especially effective in high-dimensional feature spaces.
- MLP: A feedforward neural network comprising of one hidden layer, proficient at learning intricate and nonlinear relationships among integrated features. Enhances the model's capability to distinguish subtle behavioral patterns.
- SGD: A linear classifier trained using the log function, well-suited for large-scale and sparse datasets owing to its computational efficiency and ability to converge quickly.

Each base learner generates class probabilities as intermediate outputs, forming the foundation for the next phase of classification handled by the meta-learner.

3.4.4 Meta classifier module

To improve classification performance, the system uses a stacking ensemble with logistic regression as a meta-classifier. It takes the prediction probabilities from three base classifiers—SVM, MLP, and SGD—as input features. This meta-classifier learns the best way to integrate these outputs to:

- Integrate the strengths of each base model.
- Reduce individual biases or weaknesses.
- Enhance generalization and robustness on new data.

This stacked approach provides a more balanced and accurate prediction, especially for complex multiclass driver behavior classification.

For the stacking ensemble, each base classifier (SVM, SGD, and MLP) outputs a class probability vector for every input sample rather than a discrete class label. These probability vectors are concatenated to construct a meta-feature vector of dimension $3 \times N$, where N denotes the total number of

predefined target classes.

A logistic regression meta-classifier is trained on these meta-features using a five-fold cross-validation protocol on the training set to minimize information leakage and decrease overfitting. The regularization parameter is set to C = 1.0, and the lbfgs solver is employed for stable multi-class optimization. This probabilistic stacking approach enables the meta-classifier to learn an optimal combination of the complementary decision patterns produced by the heterogeneous base classifiers, thereby improving robustness and generalization under real-world driving conditions.

3.4.5 Output prediction classes

The meta-classifier's final output identifies eight driver behavior states, divided into two groups: distraction and drowsiness. The distraction states — BottomCenter, MiddleCenter, MiddleLeft, and MiddleRight—indicate different head poses and gaze directions to assess driver attention. The states of drowsiness, closed eyes, open eyes, no yawn, and yawn reflect key physiological signs such as eye closure and yawning. Together, these predictions provide a comprehensive analysis of driver alertness and focus, enabling timely detection of risky driving behavior.

3.5 Dataset description

The study used a merged dataset of 41,094 labeled images across eight classes: four head poses (BottomCenter, MiddleCenter, MiddleLeft, MiddleRight), two eye states (Open, Closed), and two yawning statuses (yawn, no yawn). These classes highlight key visual signs of driver drowsiness and inattention. The dataset was divided into training (70%, 28,509 images), validation (15%, 6,160), and testing (15%, 6,175) sets to support effective model training and evaluation. Real-time data augmentation was applied to each subset. This diverse and well-structured dataset enables robust hybrid models, such as HDFSNet, to accurately detect and classify driver states from visual cues. To reproduce the reported performance (≈ 99.08% accuracy), a minimum of 30,000 well-balanced and diverse images with data augmentation is recommended. This ensures sufficient coverage of real-world variations in lighting conditions, facial expressions, and head poses, supporting robust generalization of the HDFSNet model in practical deployment scenarios.

Figure 2 illustrates various driver states used for training and evaluation. The top row shows head pose variations: (a) BottomCenter, (b) MiddleCenter, (c) MiddleLeft, and (d) MiddleRight. The bottom row depicts eye and yawning states: (e) Closed, (f) Open, (g) no_yawn, and (h) yawn. These categories help to assess driver drowsiness and distraction under various conditions.

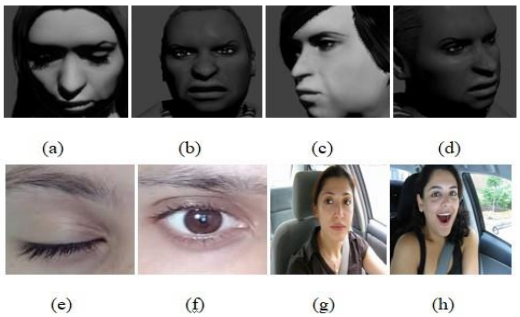

Figure 2. Typical samples from the merged dataset

The merged dataset integrates an internally curated driver monitoring dataset (head pose classes) and a public Kaggle dataset (https://www.kaggle.com/datasets/serenaraju/yawn-eye-dataset-new/data) for drowsiness-related classes. The internal dataset is not publicly distributed due to data ownership and privacy constraints. The original merged dataset contained a total of 17,789 images and exhibited significant class imbalance. To address this, offline data augmentation (±15° rotation, horizontal flipping, and brightness scaling of 0.7–1.3) was applied to underrepresented classes. After augmentation, the dataset was expanded to a total of 41,094 images, with all eight classes balanced to contain an equal number of samples. The balanced dataset was subsequently partitioned into training (70%), validation (15%), and testing (15%) subsets.

4. RESULTS AND DISCUSSION

This section thoroughly evaluates the HDFSNet architecture for driver state classification, combining hybrid deep feature extraction and ensemble learning. The experiments used a custom dataset of 41,094 images across eight classes, split into training, validation, and testing. Baseline classifiers—SVM, SGD, and MLP—were trained on features extracted from VGG16, ResNet50, and InceptionV3. Final predictions were aggregated through a logistic regression meta-classifier. The effectiveness of the proposed model was rigorously assessed using a comprehensive set of evaluation metrics, encompassing accuracy, precision, recall, F1-score, confusion matrix, ROC, and PR curves, enabling comprehensive performance comparison.

4.1 Performance of proposed Hybrid Deep Feature Fusion and Stacked Classifier Network

4.1.1 Confusion matrix analysis of HDFSNet

The confusion matrix for HDFSNet model, Figure 3, demonstrates a highly accurate classification across the eight categories of driver behavior. Key highlights include:

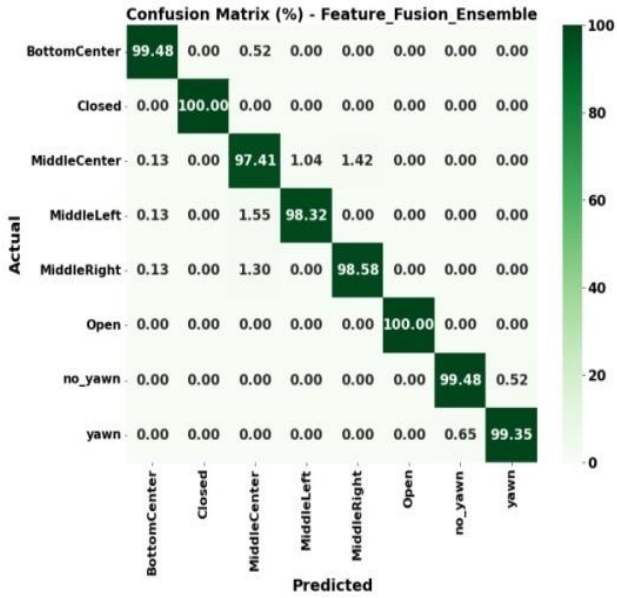

Figure 3. Confusion matrix of the proposed Hybrid Deep Feature Fusion and Stacked Classifier Network (HDFSNet) model

- The closed and open classes are classified with 100% accuracy, showing a perfect understanding of the eye states by the model.
- BottomCenter is identified with 99.48% accuracy, with a minor 0.52% misclassified as MiddleCenter.
- MiddleCenter, MiddleLeft, and MiddleRight exhibit slightly lower but still strong accuracies (97.41%, 98.32%, and 98.58%, respectively), primarily due to inter-class similarity in facial positions.
- No_yawn and yawn are predicted with 99.48% and 99.35% accuracy, respectively, which shows improved discrimination compared to previous models.

HDFSNet shows a clear performance gain compared to individual feature fusion classifiers, particularly in reducing inter-class confusion among visually similar classes (like Middle positions and yawn states). The stacked architecture with a logistic regression classifier enhances robustness and generalization, making HDFSNet the most reliable model in the proposed framework.

4.1.2 Classification report analysis of HDFSNet

Table 1, the classification report of the proposed HDFSNet architecture, demonstrates highly robust performance across all eight categories of driver behavior. The model achieved an overall accuracy of 99.08%, reflecting its robust generalization and classification capability.

Table 1. Classification report of Hybrid Deep Feature Fusion and Stacked Classifier Network (HDFSNet) performance

Class	Precision	Recall	F1-Score
BottomCenter	1.00	0.99	1.00
Closed	1.00	1.00	1.00
MiddleCenter	0.97	0.97	0.97
MiddleLeft	0.99	0.98	0.99
MiddleRight	0.99	0.99	0.99
Open	1.00	1.00	1.00
no_yawn	0.99	0.99	0.99
yawn	0.99	0.99	0.99
Accuracy			0.99
Macro Avg	0.99	0.99	0.99
Weighted Avg	0.99	0.99	0.99

HDFSNet achieved perfect F1-scores (1.00) for key classes such as Closed, Open, and BottomCenter, demonstrating flawless precision and recall. Minor dips were observed for MiddleLeft (0.99) and MiddleCenter (0.97), likely due to subtle overlaps in gaze direction. Overall, macro and weighted averages of 0.99 confirm the model's consistency across both majority and minority classes. This superior performance surpasses individual classifiers (SVM, SGD, MLP), establishing HDFSNet as a robust strategy for real-time driver behavior classification.

4.1.3 Cross-validation evaluation

To ensure robustness and reproducibility, five-fold cross-validation was performed on the combined training and validation dataset, where four folds were used for training and one fold for testing in each iteration. The mean classification accuracy and macro-averaged F1-score, along with their standard deviations, are reported to quantify performance stability across data splits.

Table 2 validates the stability of HDFSNet using 5-fold cross-validation. The model demonstrated exceptional consistency, achieving a mean accuracy of 99.08% ± 0.05, with individual folds ranging tightly between 99.00% and 99.13%. Stable precision, recall, and F1-scores of 0.99 across all folds confirm the model's robustness and effectively rule out overfitting.

4.1.4 ROC curves analysis of HDFSNet

The ROC curves of the proposed HDFSNet ensemble model, as depicted in Figure 4, demonstrate exceptional classification performance across all eight driver behaviour categories. Each class—BottomCenter, Closed, MiddleCenter, MiddleLeft, MiddleRight, Open, no yawn, and yawn—achieved an AUC of 1.00, indicating ideal sensitivity and specificity. This means that the model makes no false positive or false negative predictions for any class.

Table 2. Five-fold cross-validation performance of HDFSNet on the balanced merged dataset

Fold	Accuracy (%)	Precision	Recall	F1-Score
Fold 1	99.12	0.99	0.99	0.99
Fold 2	99.05	0.99	0.99	0.99
Fold 3	99.10	0.99	0.99	0.99
Fold 4	99.00	0.99	0.99	0.99
Fold 5	99.13	0.99	0.99	0.99
Mean ± SD	99.08 ± 0.05	0.99 ± 0.01	0.99 ± 0.01	0.99 ± 0.01

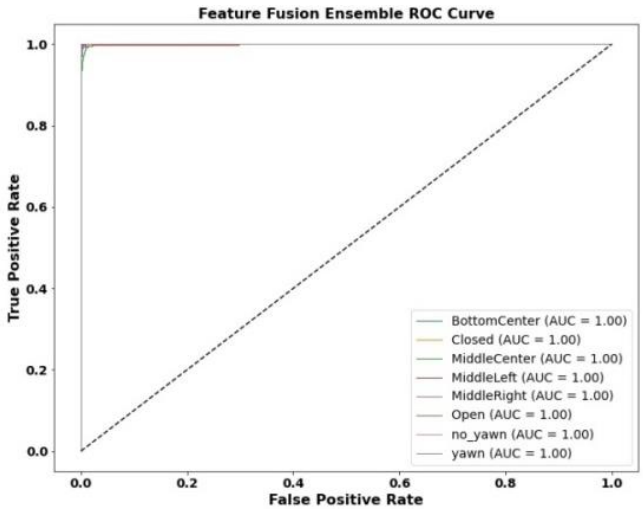

Figure 4. ROC curves for the proposed Hybrid Deep Feature Fusion and Stacked Classifier Network (HDFSNet) model

4.1.5 Precision-Recall curve analysis for HDFSNet

The PR curve depicted in Figure 5 represents the HDFSNet ensemble model's performance across all eight driver behavior classes. Each class—BottomCenter, Closed, MiddleCenter, MiddleLeft, MiddleRight, Open, no yawn, and yawn—attained a perfect AUC of 1.00.

The proposed HDFSNet framework demonstrates superior performance in detecting driver behaviors by leveraging hybrid deep feature fusion and ensemble learning. The fusion of deep features from VGG16, ResNet50, and InceptionV3 enables the model to capture diverse spatial, contextual, and abstract representations of input images. These comprehensive features are subsequently classified using three base classifiers, SVM, SGD, and MLP, whose predictions are integrated using a meta classifier to form the ultimate decision output.

The custom dataset is balanced across eight behavior classes

and contains a total of 41,094 images. The images were partitioned into training, validation, and testing subsets using a 70:15:15 split ratio to ensure consistent class representation across all phases of model development.

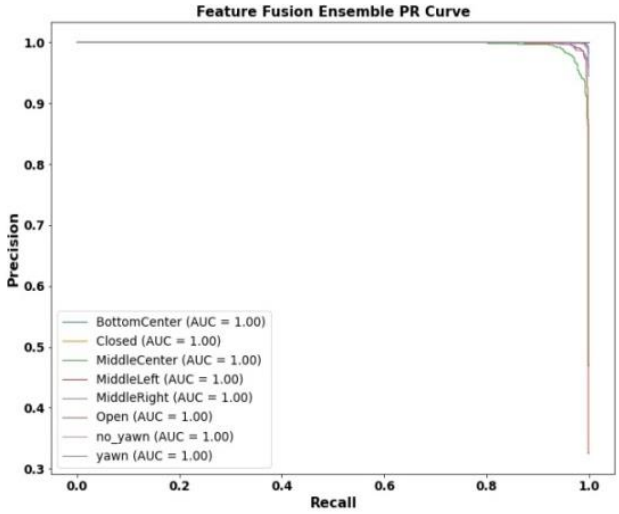

Figure 5. PR curves for the proposed Hybrid Deep Feature Fusion and Stacked Classifier Network (HDFSNet) model

At low false-positive rates critical for safety, HDFSNet performs exceptionally. All classes achieve precision and recall ≥ 0.97, with safety-critical states such as eye closure and yawning, from 0.99 to 1.00. Almost perfect AUC-PR scores (≈ 1.00) confirm highly reliable detection with minimal risk of misclassification.

4.1.6 Computational complexity and efficiency analysis

A systematic computational efficiency analysis is conducted to evaluate the performance-overhead trade-off of the HDFSNet framework against single-CNN baselines (VGG16, ResNet50, and InceptionV3 combined with SVM, SGD, or MLP classifiers) in terms of floating-point operations (FLOPs), total parameter count, memory footprint, and inference time.

Table 3. Computational complexity of feature extraction backbones

Model	GFLOPs	Parameters (Millions)	Memory Footprint (MB)
VGG16	30.71	14.70	56.13
ResNet50	7.75	23.59	89.98
InceptionV3	5.69	21.80	83.17

Table 3 shows the feature extraction backbones exhibit distinct computational profiles. VGG16 requires 30.71 GFLOPs with 14.7 million parameters (56.13 MB), ResNet50 requires 7.75 GFLOPs with 23.59 million parameters (89.98 MB), and InceptionV3 requires 5.69 GFLOPs with 21.80 million parameters (83.17 MB). By integrating these complementary models within a fused and stacked architecture, HDFSNet introduces additional computational cost relative to single-backbone pipelines.

Despite this overhead, Table 4 depicts that the average inference time remains suitable for near real-time deployment, measured at approximately 0.1939 seconds per image on the test set. This moderate increase in latency is accompanied by a consistent improvement in classification accuracy from 96.89%–98.25% for individual CNN-classifier models to 99.08% for the proposed HDFSNet framework, along with a measurable reduction in inter-class confusion among visually similar driver states.

Table 4. Performance–Efficiency comparison of Hybrid Deep Feature Fusion and Stacked Classifier Network (HDFSNet) and baseline models

Model Configuration	Feature Backbones	Classifier Strategy	Inference Time (s/image)	Accuracy (%)
Single-CNN Baseline	VGG16 / ResNet50 / InceptionV3	SVM / SGD / MLP	0.12–0.16	96.89–98.25
Proposed Model	VGG16 + ResNet50 + InceptionV3	Stacked (SVM + SGD + MLP → LR)	0.19	99.08

Table 5. Comparison of Hybrid Deep Feature Fusion and Stacked Classifier Network (HDFSNet) with state-of-the-art methods for driver distraction and drowsiness detection

Author and Year	Method	Dataset	Region of Interest	Accuracy (%)
Hssayeni et al. [11] & 2017	ResNet-152, VGG16 with transfer learning	Dashboard cam	Face images	85.00 (ResNet-152)
Omerustaoglu et al. [12] & 2020	CNN + LSTM multimodal fusion	Custom + public	Face + sensor data	85.00
Chillakuru et al. [14] & 2024	CNN+PCA+DBN+ AGBO	Public collected	Face	97.67
Lamouchi et al. [16] & 2025	LBP-TOP+SVM	UTA-RLDD, DROZY	Spatio temporal facial features	90.00
Doshi [20] & 2025	Anchor-ViT	State Farm	Upper body, Hands	92.30
Li et al. [21] & 2024	CoViT (CNN+ViT)	SFD2	body, Face	97.89
Huang et al. [22] & 2024	RFE-SHAP + Multilayer Stacking	Driver Mental Load and Emotion Dataset	Face + Physiological + Behavioral Features	80.84
Priyanka et al. [23] & 2024	Multimodal Data Fusion	Multimodal Driver Dataset	Physiological Signals + Face	96
Hybrid Model (CNN + ML Classifier) (Our Models)	VGG16, ResNet50, InceptionV3 with SVM, SGD, and MLP Classifiers	Merged Dataset	Face, eyes, mouth	96.89 to 98.25
Proposed Model	HDFSNet	Merged Dataset	Face, eyes, mouth	99.08

4.2 Model comparison

Table 5 compares several recent studies on driver state detection, showing a steady shift from traditional deep learning models toward hybrid and attention-based approaches. For example, Omerustaoglu et al. [12] combined facial images with in-vehicle sensor data using a CNN–LSTM framework and reported an accuracy of 85.00%. Similarly, Hssayeni et al. [11] applied transfer learning with deep CNN models such as ResNet-152 and achieved about 85.00% accuracy using dashboard camera images of drivers' faces. More recent hybrid methods have improved performance, with Chillakuru et al. [14] reaching 97.67% accuracy by integrating CNN features with PCA and deep belief networks. Transformer-based models have also gained attention. Doshi [20] introduced Anchor-ViT, which focuses on upper-body and hand regions and achieved 92.30% accuracy on the State Farm dataset, while Li et al. [21] proposed a CNN–ViT hybrid model (CoViT) that achieved 97.89% accuracy on the SFD2 dataset by combining body and facial information. Traditional spatiotemporal approaches, such as the LBP-TOP and SVM method by Lamouchi et al. [16], reported 90.00% accuracy on the UTA-RLDD and DROZY datasets. Huang et al. [22] further explored a multilayer stacking approach with RFE-SHAP feature selection and achieved 97.48% accuracy in recognizing drivers' mental and emotional states. In comparison, the hybrid CNN–ML baseline models in this study achieved accuracies between 96.89% and 98.25% on the merged facial dataset. The proposed HDFSNet model outperformed these approaches with an accuracy of 99.08%, highlighting the benefit of combining multi-CNN feature fusion with a stacked classification strategy for more reliable and robust driver behavior recognition.

These results empirically confirm the limitations of single-backbone architectures in capturing subtle micro-expressions and marginal head pose variations and directly justify the necessity of the proposed multi-CNN feature fusion and stacked classification framework, as theoretically motivated in the Introduction.

An ablation study was performed to determine the contributions of feature fusion and classifier stacking in HDFSNet. Two configurations were evaluated: (i) single-CNN pipelines with individual classifiers (VGG16, ResNet50, or InceptionV3 combined with SVM, SGD, or MLP), and (ii) the full HDFSNet framework with multi-CNN feature fusion and logistic regression-based stacking. The results show a clear performance gain, with single-CNN models achieving accuracies of 96.89%–98.25%, while HDFSNet attains the highest accuracy of 99.08%, confirming improved generalization and reduced inter-class confusion for subtle driver behavior recognition.

5. CONCLUSION

This study presents an advanced and efficacious approach to driver distraction and drowsiness detection through the proposed HDFSNet model, which leverages hybrid deep feature fusion and stacked classifiers. By combining the strengths of multiple pre-trained CNN architectures and integrating them with SVM, SGD, and MLP classifiers, the model achieves a superior accuracy of 99.08% on merged datasets. Comparative analysis with existing methods highlights its robustness and efficiency in identifying driver fatigue and inattention. The findings indicate that the suggested framework not only outperforms individual and fusion-based models but also offers a scalable solution for enhancing road safety.

This study advances driver behavior recognition by addressing three critical gaps in existing systems: limited feature representation, ineffective feature integration, and poor generalization under real-world variability. The proposed HDFSNet framework demonstrates that multi-level feature fusion across complementary CNN backbones enables the simultaneous modeling of fine-grained facial micro-expressions and global head pose context, providing a richer and more discriminative representation than single-backbone approaches. Beyond representation, this work establishes the role of decision-level classifier stacking as a generalization mechanism, where heterogeneous base classifiers learn complementary decision boundaries and a logistic regression meta-classifier optimally integrates their probabilistic outputs. This strategy consistently reduces inter-class confusion among visually similar driver states and enhances robustness to variations in illumination, occlusion, and pose. Together, these contributions provide both a theoretical and applied justification for hybrid feature fusion and stacked ensemble learning in safety-critical driver monitoring systems, demonstrating that the performance gains of HDFSNet stem from principled architectural design rather than increased model complexity alone.

Future work will focus on improving the model by incorporating larger and more diverse datasets, integrating physiological signals (EEG, heart rate, etc.), and optimizing it for real-time deployment on embedded systems. Additionally, incorporating temporal and multimodal learning techniques could further boost its reliability in real-world driving environments.

ACKNOWLEDGEMENTS

The authors sincerely thank VIT-AP University for providing the necessary research support, infrastructure, and resources for this work.

REFERENCES

[1] Almazroi, A.A., Alqarni, M.A., Aslam, N., Shah, R.A. (2023). Real-time CNN-based driver distraction & drowsiness detection system. Intelligent Automation & Soft Computing, 37(2): 2153-2174. https://doi.org/10.32604/iasc.2023.039732

[2] Venkateswarlu, M., Ch, V.R.R. (2024). DrowsyDetectNet: Driver drowsiness detection using lightweight CNN with limited training data. IEEE Access, 12: 110476-110491. https://doi.org/10.1109/ACCESS.2024.3440585

[3] Streiffer, C., Raghavendra, R., Benson, T., Srivatsa, M. (2017). DaRNet: A deep learning solution for distracted driving detection. In Proceedings of the 18th ACM/IFIP/USENIX Middleware Conference: Industrial Track, pp. 22-28. https://doi.org/10.1145/3154448.3154452

[4] Ping, P., Huang, C., Ding, W., Liu, Y., Chiyomi, M., Kazuya, T. (2023). Distracted driving detection based on the fusion of deep learning and causal reasoning.

Information Fusion, 89: 121-142. https://doi.org/10.1016/j.inffus.2022.08.009
[5] Ezzouhri, A., Charouh, Z., Ghogho, M., Guennoun, Z. (2021). Robust deep learning-based driver distraction detection and classification. IEEE Access, 9: 168080-168092. https://doi.org/10.1109/ACCESS.2021.313379
[6] National Highway Traffic Safety Administration (NHTSA). Distracted Driving. https://www.nhtsa.gov/risky-driving/distracted-driving, accessed on Oct. 26, 2025.
[7] Bentamou, A., Chretien, S., Gavet, Y. (2025). 3D denoising diffusion probabilistic models for 3D microstructure image generation of fuel cell electrodes. Computational Materials Science, 248: 113596. https://doi.org/10.1016/j.commatsci.2024.113596
[8] National Highway Traffic Safety Administration (NHTSA). Drowsy Driving. https://www.nhtsa.gov/risky-driving/drowsy-driving, accessed on Apr. 26, 2025.
[9] Dua, M., Shakshi, Singla, R., Raj, S., Jangra, A. (2021). Deep CNN models-based ensemble approach to driver drowsiness detection. Neural Computing and Applications, 33(8): 3155-3168. https://doi.org/10.1007/s00521-020-05209-7
[10] Venkateswarlu, M., Chirra, V.R.R. (2025). CNN-ViT: A multi-feature learning based approach for driver drowsiness detection. Array, 27: 100425. https://doi.org/10.1016/j.array.2025.100425
[11] Hssayeni, M.D., Saxena, S., Ptucha, R., Savakis, A. (2017). Distracted driver detection: Deep learning vs handcrafted features. Electronic Imaging, 29: 20-26. https://doi.org/10.2352/ISSN.2470-1173.2017.10.IMAWM-162
[12] Omerustaoglu, F., Sakar, C.O., Kar, G. (2020). Distracted driver detection by combining in-vehicle and image data using deep learning. Applied Soft Computing, 96: 106657. https://doi.org/10.1016/j.asoc.2020.106657
[13] Reddy, C.V.R., Reddy, U.S., Babu, D.M. (2019). An automatic driver drowsiness detection system using DWT and RBFNN. International Journal of Recent Technology and Engineering (IJRTE), 7(5S4): 41-44.
[14] Chillakuru, P., Ananthajothi, K., Divya, D. (2024). Three stage classification framework with ranking scheme for distracted driver detection using heuristic-assisted strategy. Knowledge-Based Systems, 293: 111589. https://doi.org/10.1016/j.knosys.2024.111589
[15] Hasan, M.M., Watling, C.N., Larue, G.S. (2024). Validation and interpretation of a multimodal drowsiness detection system using explainable machine learning. Computer Methods and Programs in Biomedicine, 243: 107925. https://doi.org/10.1016/j.cmpb.2023.107925
[16] Lamouchi, D., Yaddaden, Y., Parent, J., Cherif, R. (2025). Efficient driver drowsiness detection using spatiotemporal features with support vector machine. International Journal of Intelligent Transportation Systems Research, (in press), 1-13.
[17] Simonyan, K., Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556. https://doi.org/10.48550/arXiv.1409.1556
[18] He, K., Zhang, X., Ren, S., Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, pp. 770-778. https://doi.org/10.1109/CVPR.2016.90
[19] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, pp. 2818-2826. https://doi.org/10.1109/CVPR.2016.308
[20] Doshi, V. (2025). Anchor-ViT: Spatially-focused vision transformer for distracted driving detection. In 2025 IEEE International Conference on Image Processing (ICIP), Anchorage, AK, USA, pp. 2700-2705. https://doi.org/10.1109/ICIP55913.2025.11084655
[21] Li, Z., Zhao, X., Wu, F., Chen, D., Wang, C. (2024). A lightweight and efficient distracted driver detection model fusing convolutional neural network and vision transformer. IEEE Transactions on Intelligent Transportation Systems, 25(12): 19962-19978. https://doi.org/10.1109/TITS.2024.3447041
[22] Huang, J., Peng, Y., Hu, L. (2024). A multilayer stacking method based on RFE-SHAP feature selection strategy for recognition of driver's mental load and emotional state. Expert Systems with Applications, 238: 121729. https://doi.org/10.1016/j.eswa.2023.121729
[23] Priyanka, S., Shanthi, S., Kumar, A.S., Praveen, V. (2024). Data fusion for driver drowsiness recognition: A multimodal perspective. Egyptian Informatics Journal, 27: 100529. https://doi.org/10.1016/j.eij.2023.100529