Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Research on a Safe Driving Monitoring System Based on Multimodal Data

hao Cui
China Utone Construction Consulting Co., Ltd
Jie Cai
China Utone Construction Consulting Co., Ltd
Lingling Tian
China Utone Construction Consulting Co., Ltd
Zepeng Wei
1223076601@njupt.edu.cn

Nanjing University of Posts and Telecommunications
Dengyin Zhang
Nanjing University of Posts and Telecommunications

Research Article

Keywords: Emotional Driving Detection, Fatigued Driving Detection, Distracted Driving Detection, Multimodal Sentiment Analysis, Monitoring of Construction Machinery Driving Behavior

Posted Date: November 17th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-8619471/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Research on a Safe Driving Monitoring System Based on Multimodal Data

Hao Cui[1†], Jie Cai[1†],Lingling Tian[1†], Zepeng Wei[2*] and Dengyin Zhang[2*]

[1]China Post Construction Consulting Co. Ltd, Nanjing 210003,Jiangsu,China

[2]School of Internet of Things, Nanjing University of Posts and Telecommunications, Nanjing 210003,Jiangsu,China

[*]Corresponding authors. E-mails:1223076601@njupt.edu.cn, zhangdy@njupt.edu.cn

Contributing authors: cuihao.cutcc@chinaccs.cn, caijie.cutcc@chinaccs.cn, tianlingling.cutcc@chinaccs.cn

[†]These authors contributed equally to this work.

Abstract

To address the safety issues in construction machinery caused by abnormal driving behaviors, this paper proposes a Multimodal Safe Driving Monitoring System (MSDMS). The system consists of two main components. The first is the Multimodal Sentiment Analysis model based on Hybrid Attention (MSA-HA), designed for emotional driving detection. By integrating hybrid attention mechanisms with a Tensor Fusion Network, this model enhances multimodal feature interaction and improves the accuracy of emotion detection. The second component employs the YOLOv8n-DPCLN model for the detection of fatigue and distracted driving. In this model, a Deepwise Partial Convolution Layer Network (DPCLN) replaces the bottleneck within the C2f module to enhance overall performance. Experimental results demonstrate that the MSA-HA achieves binary and ternary classification accuracies of 83.5% and 74.76%, respectively, on the SIMSv2 dataset, outperforming other mainstream models. Compared to YOLOv8n, the YOLOv8n-DPCLN improves mAP@0.5 and mAP@0.5-0.95 by 1.9% and 3.7%, respectively, while reducing the parameter count by 0.4 M and computational cost by 1.1 GFLOPs.Finally, MSDMS achieves recognition rates of 78.7%, 93%, and 86% for emotion, fatigue, and distracted driving, respectively, thereby satisfying the requirements for safe driving monitoring.

Keywords: Emotional Driving Detection, Fatigued Driving Detection, Distracted Driving Detection, Multimodal Sentiment Analysis, Monitoring of Construction Machinery Driving Behavior

1 Introduction

With the vigorous development of the construction machinery industry, equipment operational safety has garnered increasing attention. Research indicates that human operational factors are one of the primary causes of safety accidents[1]. The extremely harsh construction environment imposes a severe physical and mental burden on drivers, often leaving them in states of fatigue, distraction, and emotional instability[2]. Consequently, the development of intelligent safety systems capable of real-time, accurate monitoring of driver status and providing early warnings has become a key technological means to ensure the operational safety of construction machinery. Early driver monitoring systems relied on vehicle sensors or invasive physiological signals (such as electroencephalograms，electro-oculogram, electromyogram and electrocardiogram) for state inference, but they commonly suffered from issues like insufficient accuracy, high cost, or difficult

deployment. With the rapid development of deep learning and computer vision technologies, data based on images and audio can provide a more comprehensive assessment of the driver's real-time status. Abbas et al. proposed a driver distraction and fatigue detection system based on multi-view cameras and biosignal sensors. They utilized Convolutional Neural Networks (CNNs) and Long Short-Term Memory network (LSTM) to extract physiological features and a Deep Residual Neural Network (DRNN) to fuse these features for distraction and fatigue detection[3]. Although this method achieved certain results, it requires information from electrocardiograms, electroencephalograms, myoelectricity, and electrooculograms, leading to high equipment deployment costs. Wang et al. proposed a driving status monitoring system combining YOLOv7 and Dlib for face detection. YOLOv7 was used for distraction detection, while Dlib was used for face detection, with a CNN extracting key facial features to detect driver fatigue, achieving a high recognition rate[4]. Dong et al. used a Single Shot Scale-invariant Face Detector (S3FD) to detect faces in images, then employed a Face Alignment Network (FAN) to extract facial features. They utilized random forest techniques to analyze fatigue driving states and subsequently used a CNN for distraction detection, achieving high detection accuracy[5]. Although the aforementioned systems have integrated fatigue and distraction detection and achieved certain success, the emotional state of the driver is equally crucial in monitoring driving behavior. Emotional abnormalities, such as negative emotions like anxiety, anger, or depression, can significantly impair a driver's attention and decision-making abilities, thereby increasing the risk of accidents[6]. However, current comprehensive monitoring systems have not yet incorporated emotional driving detection into the overall monitoring framework, leaving a blind spot in the system's ability to recognize the driver's state. Therefore, constructing a comprehensive driver monitoring system that integrates fatigue, distraction, and emotional driving detection can more holistically assess the driver's condition, effectively improving the accuracy and practicality of safety monitoring for construction machinery.

In the context of emotional driving detection, current research predominantly employs multimodal fusion methods. Ying et al. developed a multimodal sentiment analysis model by establishing a facial expression recognition network and a lightweight speech emotion recognition network, combining audio and image features, and achieved good detection results in emotional driving detection tasks[7]. Although this model effectively fuses audio and visual modalities, it does not consider the important role of textual information in sentiment analysis, which may lead to insufficient recognition capabilities for certain emotional states. Sun et al. conducted feature relationship analysis and feature extraction on drivers' electrocardiogram signals and driving behavior signals, and used the SVM algorithm to build a binary recognition model for anger, which effectively improved the recognition accuracy of driver anger[8]. However, the electrocardiogram data required by this method needs professional equipment for accurate extraction, and its integration into vehicles would increase costs and affect driver comfort, making the technology difficult to apply.

To address the shortcomings of current driver monitoring systems in terms of comprehensiveness and practicality, this paper proposes a Multimodal Safe Driving Monitoring System (MSDMS) based on multimodal data, which can simultaneously detect three key driving states: fatigue, distraction, and emotion. The system integrates two core modules: a novel Multimodal Sentiment Analysis Model based on Hybrid Attention (MSA-HA) for emotional driving detection, and a YOLOv8n-DPCLN model for fatigue and distraction detection. By fusing speech, text, and visual information for in-depth sentiment analysis and combining it with real-time capture of driver behavior, MSDMS can more comprehensively and accurately assess the driver's overall state. This system aims to fill the gaps in existing technology for driving behavior

detection, providing new ideas and practical solutions for building a more reliable and intelligent driving safety assurance system.

2 Multimodal Safe Driving Monitoring System

Figure 1 illustrates the architecture of the Multimodal Safe Driving Monitoring System (MSDMS). The system is primarily composed of two detection models: the Multimodal Sentiment Analysis model (MSA-HA) for the emotional driving detection task, and the YOLOv8n-DPCLN object detection model for the fatigue and distracted driving detection tasks.The system processes video information collected by a camera, separating it into dynamic image data and audio data. For the emotional driving detection task, facial features are extracted using the OpenFace toolkit, which includes 68 facial landmarks and 17 Facial Action Units[9]. The openSMILE toolkit is employed to extract eGeMAPS audio features[10,11]. Concurrently, the Vosk speech-to-text toolkit converts audio information into text, from which textual features are then extracted using the pre-trained BERT model[12]. Subsequently, the extracted initial facial, audio, and textual features are fed into the MSA-HA model for emotion recognition. An alert is triggered if negative emotions are detected. For the fatigue and distracted driving detection tasks, the YOLOv8n-DPCLN object detection model processes the dynamic images to identify whether the driver is in a state of fatigue or distraction.

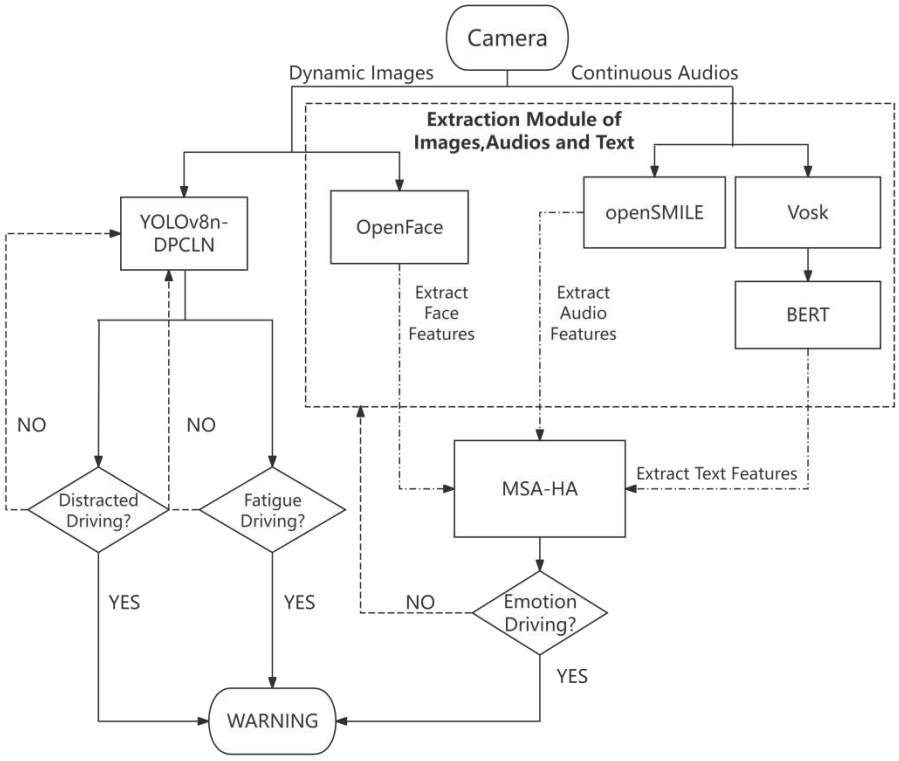


Fig.1 Multimodal Safe Driving Monitoring System

2.1 Multimodal sentiment analysis model

To address the data selection deficiencies in current emotional driving detection, this paper utilizes multimodal data that is readily accessible and highly correlated: text, audio, and visual information. Accordingly, we propose a Multimodal Sentiment Analysis model based on a Hybrid Attention mechanism

(MSA-HA), the architecture of which is depicted in Figure 2.In this model, audio and visual features are extracted using an mLSTM network[13]. mLSTM extends the vector operations of traditional LSTM to matrix operations, significantly enhancing the model's memory capacity and parallel processing capabilities. This allows it to capture longer dependencies and more complex input features during training . As shown in Figure 2, the initial audio and visual features are processed by the mLSTM to yield their respective feature outputs and final hidden states. For the initial textual features, since they have already been processed by the pre-trained BERT model to encapsulate important features such as contextual information, syntactic structure, and semantics, they are processed directly using a linear layer. These unimodal features are then passed through a hybrid attention layer, which is composed of a multi-head attention layer and a cross-attention layer, to output cross-modal features.

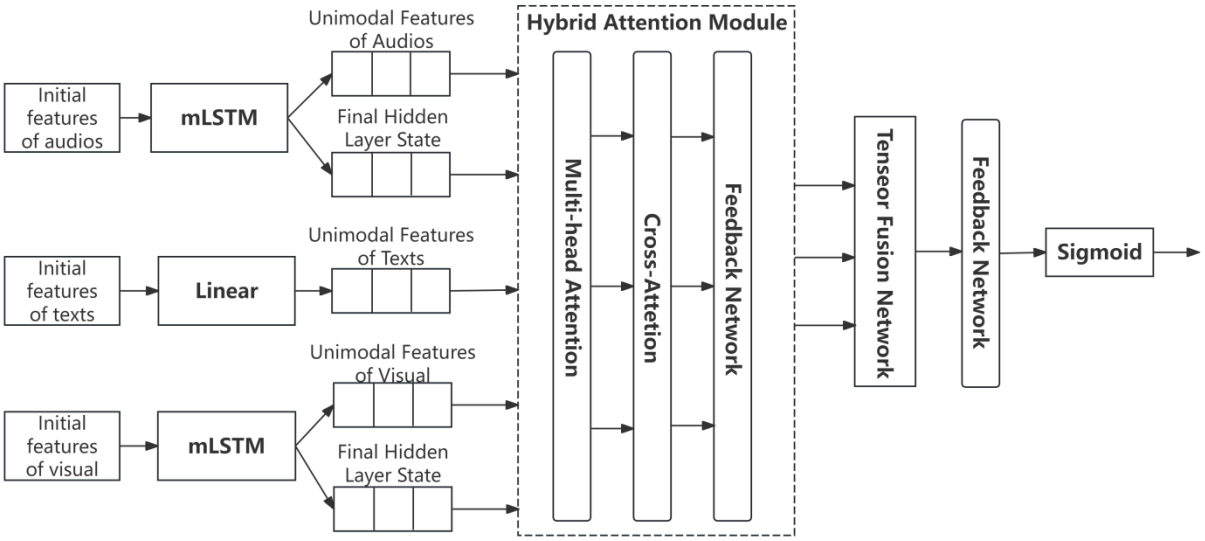


Fig.2 Multimodal Sentiment Analysis model based on Hybrid Attention

The multi-head attention layer is used to learn features of different modalities from different subspaces, and its parallel structure can enhance the expressive power of unimodal features [14]. In the multi-head attention layer, the feature representation of the visual modality can be obtained by the following formula:

$$headi = softmax(\frac{Q_v W_i^Q (K_v W_i^K)^T}{\sqrt{d_k}}) V_v W_i^v \tag{1}$$

$$Multihead(Q_v, K_v, V_v) = Concat(head_i, head_2, ..., head_n) W^O \tag{2}$$

Wherein, Q_v , K_v and V_v represent the query , key and value matrix,respectively,which are obtained from the input visual features through linear mapping. W_i^Q , W_i^K and W_i^V denote the corresponding weight matrices for Q_v , K_v and V_v ,while W^O represents the weight matrix for the output. d_k is the dimension of K_v . $\sqrt{d_k}$ is the scaling factor, which prevents the output of the *softmax* activation function from becoming too extreme by reducing the dot product values, thereby improving training stability. Similarly, the audio modalities' feature M_A and text modalities' feature M_T can be obtained by the following formulas:

$$M_A = Multihead(Q_A, K_A, V_A) \tag{3}$$

$$M_T = Multihead(Q_T, K_T, V_T) \tag{4}$$

The output of the multi-head attention layer is then fed into a cross-attention layer to obtain cross-modal

feature representations. The cross-attention layer effectively strengthens the connections between different modalities while capturing the mixed features of text and speech, text and facial expressions, and speech and facial expressions. This allows for a more accurate analysis and prediction of emotional states. Taking audio and text features as an example, Figure 3 illustrates the structure of the hybrid attention mechanism.

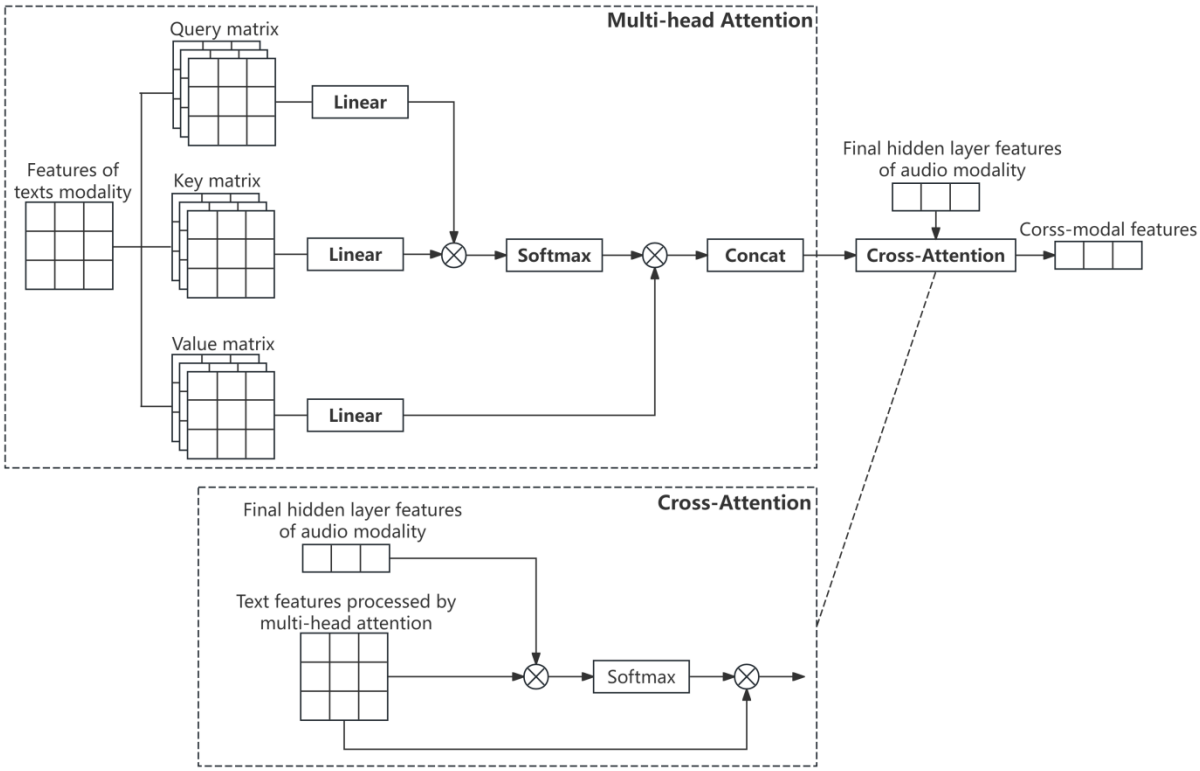


Fig.3 Hybrid attention mechanism

The cross-modal features F_{AT} between the audio and text modalities can be obtained by the following formula:

$$F_{AT} = \text{softmax}(\frac{M_T h_A}{\sqrt{\text{d}_{h_A}}})M_T \tag{5}$$

Wherein, h_A represents the final hidden layer features of the initial audio features after being processed by the mLSTM network, and M_T represents the features of the text modality after being processed by the multi-head attention mechanism. the cross-modal features between the audio and visual modalities, F_{AV}, and the cross-modal features between the visual and text modalities, F_{VT}, can be obtained through the following formulas, respectively:

$$F_{AV} = \text{softmax}(\frac{M_V h_A}{\sqrt{d_{h_A}}})M_V \tag{6}$$

$$F_{VT} = \text{softmax}(\frac{M_T h_V}{\sqrt{\text{d}_{h_V}}})M_T \tag{7}$$

Subsequently, these cross-modal features are passed through a feed-forward neural network, which consists of two fully-connected layers and a ReLU activation function. They are then jointly sent to a Tensor Fusion

Network for feature fusion[15]. The Tensor Fusion Network models the unimodal, bimodal, and trimodal interactions using a 3-fold Cartesian product of the modality embeddings.The specific implementation is given by the following formula:

$$Z_{bi}(Z^a, Z^b) = \begin{bmatrix} Z^a \\ 1 \end{bmatrix} \otimes \begin{bmatrix} Z^b \\ 1 \end{bmatrix} \tag{8}$$

$$Z_{tr}(Z^a, Z^b, Z^c) = \begin{bmatrix} Z^a \\ 1 \end{bmatrix} \otimes \begin{bmatrix} Z^b \\ 1 \end{bmatrix} \otimes \begin{bmatrix} Z^c \\ 1 \end{bmatrix} \tag{9}$$

Wherein, Z_{bi} represents the feature correlation between two modalities, and Z_{tr} represents the feature correlation among the three modalities. Therefore, the tensor-fused feature representation of the three cross-modal features is given by the following formula:

$$Z = Z_{tr}(F_{AT}, F_{AV}, F_{VT}) \tag{10}$$

Finally, the emotion prediction results for the multi-modal features are obtained by passing them through a feed-forward neural network, which consists of three fully-connected layers and two ReLU layers, followed by a Sigmoid activation function:

$$\text{Output_M} = \text{Sigmoid}(f(Z)) \tag{11}$$

Among them, $f()$ represents a feedforward neural network composed of 3 full connection layers and 2 ReLU layers.

The prediction results for each unimodal feature can be obtained by the following formula:

$$\text{Output_A} = f(A_f) \tag{12}$$

$$\text{Output_T} = f(T_f) \tag{13}$$

$$\text{Output_V} = f(V_f) \tag{14}$$

Wherein, A_f represents the feature output obtained after processing the initial audio features through the mLSTM network, V_f represents the feature output obtained after processing the visual features through the mLSTM network, and T_f is the text feature obtained after passing the initial text features through a linear layer.

2.2 Fatigue and distracted driving detection model

This study employs the YOLOv8n-DPCLN model, an object detection model based on YOLOv8n, to detect driver fatigue and distraction. Compared to previous YOLO versions, YOLOv8 has achieved significant improvements in both accuracy and efficiency. By optimizing the network architecture, introducing mechanisms for hybrid convolution and attention, adopting an adaptive anchor strategy, and improving the loss function, YOLOv8 enables the model to perform real-time object detection tasks more accurately and rapidly. It utilizes a state-of-the-art convolutional neural network architecture and incorporates multi-scale training and inference techniques, enabling better handling of objects of varying sizes and shapes. Simultaneously, YOLOv8 employs techniques such as data augmentation and knowledge distillation during training to further enhance the model's generalization ability. It also supports quantization and pruning, allowing for efficient operation on edge devices. By detecting in real-time information such as the closure state of the driver's mouth and eyes, as well as surrounding objects, YOLOv8 can effectively identify signs of fatigue or distracted driving, such as drowsiness or behaviors like using a mobile phone or smoking. These tasks typically demand

high-precision object detection and rapid response, requirements that YOLOv8 meets. With its high inference speed and excellent detection capabilities, it can promptly identify and provide feedback on the driver's state in real-time monitoring systems. Therefore, YOLOv8 holds significant application value for the warning and intervention of fatigue and distraction in the context of construction machinery operation detection.

Figure 4 shows the model architecture of YOLOv8. An input image first passes through a backbone for multi-scale feature extraction. Its core Spatial Pyramid Pooling Fast (SPPF) module utilizes a structure of three sequential max-pooling layers of the same kernel size, which enhances the gradient path to significantly reduce computational cost (FLOPs) and accelerate processing features while fully preserving the original receptive field. The resulting features are then fed into the Neck, which has a Path Aggregation Network with Feature Pyramid Network (PAN-FPN) structure, for multi-scale fusion. This deeply fuses high-level semantic information with low-level localization details, effectively improving the perceptual robustness for multi-scale targets. Finally, the fused features are input to the detection Head to perform object classification and bounding box regression tasks, outputting detection results that include class, confidence score, and location.

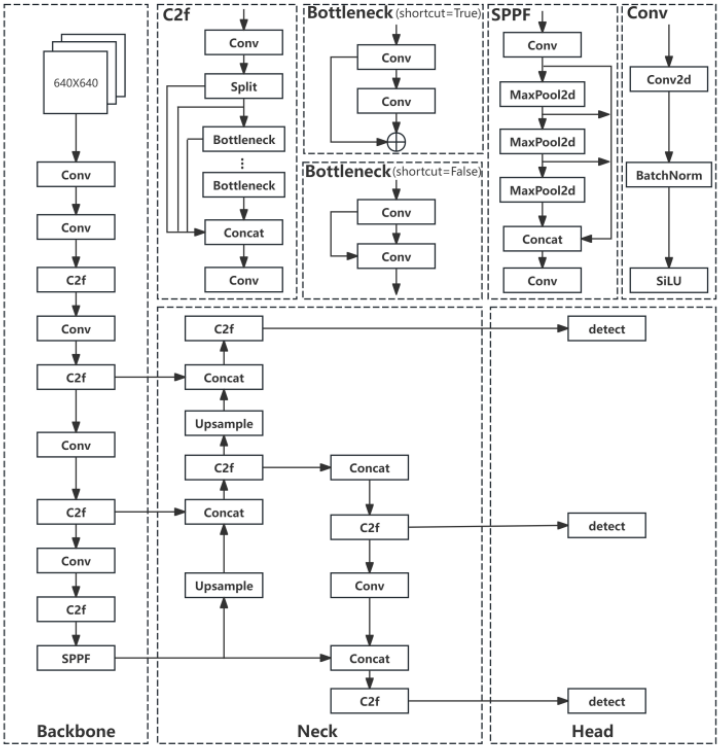


Fig.4 Structure of YOLOv8 model

YOLOv8n-DPCLN is an improved model based on YOLOv8n, with its primary structural modifications occurring in the C2f module. Specifically, the C2f module is enhanced to C2f-DPCLN by replacing thebottleneckwithin the C2f module with a Depthwise Partial Convolutional Layer Network (DPCLN), the structure of which is illustrated in Figure 5. The C2f_DPCLN module leverages an efficient partial convolution method to extract features at different scales from the input. Instead of performing convolution across all channels, it applies the operation to only a subset of channels. This, combined with Depthwise Convolution[16], further reduces the parameter count and significantly enhances computational efficiency. Finally, the model employs a 1×1 convolutional layer to fuse the features from various scales. A residual connection is then used to add the original input features to the processed features. This approach preserves the original information while incorporating new multi-scale information, thereby enhancing the model's

expressive power and performance.

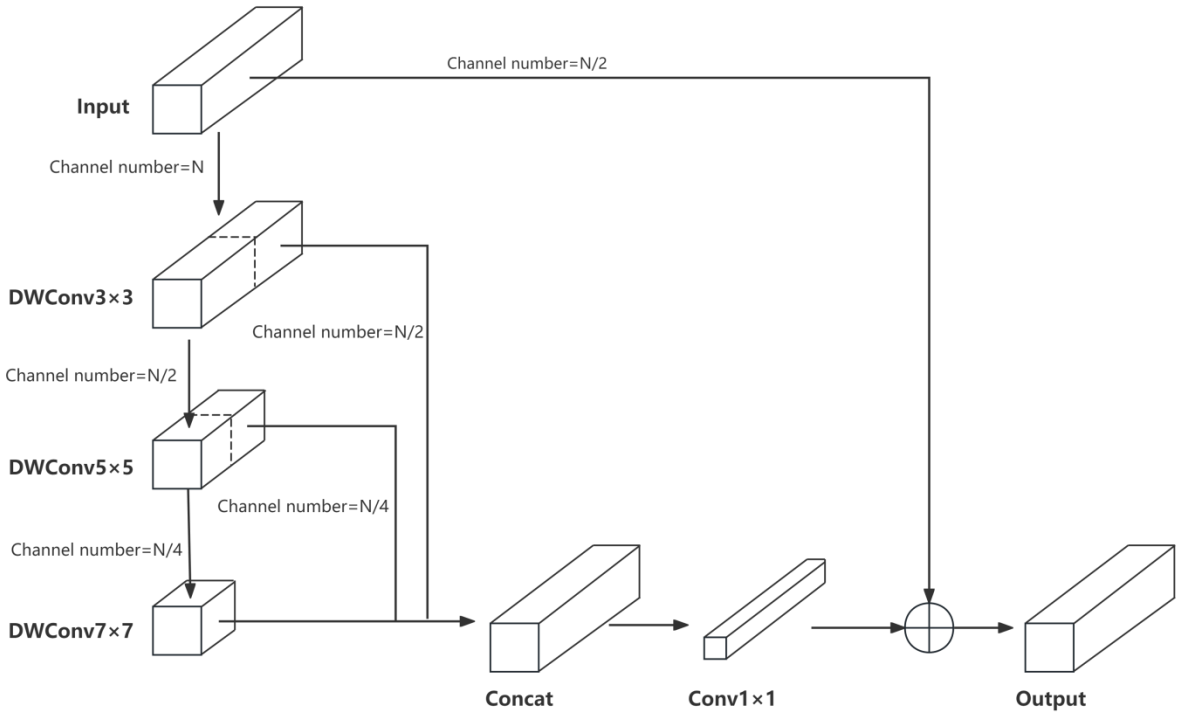


Fig.5 Depthwise partial convolutional layer network DPCLN

Let X be the input feature, and C be the number of channels. First, DPCLN processes the input using a 3×3 depthwise convolution module (DWConv3×3) to obtain the output feature X_1:

$$X_1 = \mathrm{DWConv}_{3\times3}(X) \tag{15}$$

The output of the C/2 channel in front of X_1 is sent to DWConv5×5 to obtain the output X_2:

$$X_2 = \mathrm{DWConv}_{5\times5}(X_1^{\frac{C}{2}}) \tag{16}$$

Wherein, $X_1^{\frac{C}{2}}$ represents the output of the former C/2 channels of X_1. Similarly, the output of the first C/2 channels of X_2 is fed into a DWConv7×7 to obtain the output X_3:

$$X_3 = \mathrm{DWConv}_{7\times7}(X_2^{\frac{C}{2}}) \tag{17}$$

Wherein, $X_2^{\frac{C}{2}}$ represents the output of the former C/2 channels of X_2. After obtaining the multi-scale feature map, DPCLN integrates the output of X_3 with the latter C/2 channels of X_1 and X_2 through the Concat module to produce the output X_4.

$$X_4 = \mathrm{Concat}(X_3, \hat{X}_1^{\frac{C}{2}}, \hat{X}_2^{\frac{C}{2}}) \tag{18}$$

Wherein, $\hat{X}_1^{\frac{C}{2}}$ and $\hat{X}_2^{\frac{C}{2}}$ respectively denote the outputs from the latter C/2 channels of X_1 and X_2 . This operation fuses information from different scales, ensuring the model can learn multi-scale features. Finally,

the output X_4 is processed by a 1×1 convolution and is then added to the original input feature X via a residual connection to yield the final ouput, X_out:

$$X_out = \mathrm{Add}(\mathrm{Conv}_{1\times1}(X_4), X) \tag{19}$$

The use of residual connections helps to prevent information loss while simultaneously introducing multi-scale contextual information, thereby enhancing the model's expressive power.

3 Experiments and analysis

3.1 Environment configuration, dataset and training parameters

In this study, the MSA-HA model, a multi-modal sentiment analysis model based on a hybrid attention mechanism, is employed for the detection of emotional driving. For the detection of drowsy and distracted driving, the YOLOv8n-DPCLN model is utilized. The dataset for the emotional driving detection task is the public multi-modal sentiment analysis dataset SIMSv2, with sentiment categorized into three classes: negative, neutral, and positive[17]. The training parameters are as follows: a batch size of 128, an early stopping patience of 8, and a learning rate of 0.001. The dataset for the drowsy and distracted driving detection task was created by collecting images from the Roboflow website and annotating them using the labelimg tool. This dataset comprises 1500 images, covering four types of abnormal driving behaviors (drowsy, yawning, using a phone, and smoking), with the corresponding labels “drows”, “yawning”, “phone”, and “smoking”. The dataset was partitioned into training, validation, and test sets at a 7:2:1 ratio. The sample sizes for the training, validation, and test sets are 1050, 300, and 150, respectively. The training parameters are set as follows: a batch size of 16, 300 epochs, a patience value of 50, and a learning rate of 0.01. The input image size is 640x640. The experimental environment for this study is configured as follows: the graphics card is an NVIDIA RTX 2060S (8G), the CPU is an Intel Core i5-10400F, the operating system is Windows 10, the development environment is Python 3.10, and it utilizes CUDA 12.1 with the PyTorch 2.4.1 deep learning framework.

3.2 Multimodal sentiment analysis model

3.2.1 Evaluation indicators

For the emotional driving detection task, the evaluation metrics adopted in this study are binary classification accuracy (ACC_2), three-class classification accuracy (ACC_3), F1 score, Mean Absolute Error (MAE), and the Correlation Coefficient(Corr). ACC_2 is the classification accuracy metric for negative versus non-negative emotions, and its calculation formula is as follows:

$$\mathrm{ACC_2} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \tag{20}$$

Wherein, TN represents the samples with real negative class and detected as negative class.TP represents the number of samples detected correctly in the positive sample. FP represents the number of samples detected

incorrectly in the positive sample. FN represents the number of samples with detection errors in the negative sample.

ACC_3 is an indicator of the classification accuracy of negative, neutral and positive emotions. Its calculation formula is:

$$\mathrm{ACC_3} = \frac{\mathrm{TP}_1 + \mathrm{TP}_2 + \mathrm{TP}_3}{\mathrm{N}} \tag{21}$$

Wherein, TP_1, TP_2 and TP_3 respectively represent the number of samples correctly predicted for each category (negative, neutral and positive), that is, the number of samples correctly predicted by the model for each category. N is the total number of samples, that is, the total number of all test samples.

F1 score is an important indicator used in machine learning and statistics to evaluate model performance. It combines the advantages of accuracy and recall rate, and can measure the accuracy and integrity of a model in positive prediction at the same time. Its calculation formula is:

$$\mathrm{F1} = 2 \times \frac{\mathrm{P} \times \mathrm{R}}{\mathrm{P} + \mathrm{R}} \tag{22}$$

Wherein, P is the accuracy rate and R is the recall rate.

MAE is a commonly used regression model evaluation index, used to measure the error between predicted values and true values. It is the average of the absolute value of all prediction errors, which can intuitively reflect the accuracy of model prediction. Its calculation formula is:

$$\mathrm{MAE} = \frac{1}{\mathrm{n}} \sum_{i=1}^{1} \left| y_i - \hat{y}_i \right| \tag{23}$$

Where n is the number of samples, y_i represents the true value of the i-th sample, and $\hat{y}_i$ represents the predicted value of the i-th sample.

Corr is a statistical measure that quantifies the strength and direction of a relationship between two variables. It indicates the degree of linear correlation between two variables, and is commonly used to evaluate the correlation between predicted values in regression models and their actual values. The formula for calculating it is:

$$\mathrm{Corr} = \frac{\sum_{i=1}^{\mathrm{n}} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{\mathrm{n}} (x_i - \bar{x})^2 \sum_{i=1}^{\mathrm{n}} (y_i - \bar{y})^2}} \tag{24}$$

Where x_i and y_i are the i-th data points of two variables. $\bar{x}$ and $\bar{y}$ are the means of x and y. n is the number of data points.

3.2.2 Comparative experiments

To verify the performance of the MSA-HA model, comparative experiments were conducted against other mainstream multimodal sentiment analysis models including MLF_DNN[18], MLMF[19], SELF_MM[20], BERT-MAG[21], MMIM[22], TETFN[23] and CENET[24].To ensure fairness, the environmental configurations, datasets, and training parameters employed for the comparison models were identical to those used for the MSA-HA model. Table 1 presents the results of these comparative experiments. Table 2 illustrates

the performance comparison between unimodal and multimodal detection using the MSA-HA model.

As shown in Table 1, the multi-modal sentiment analysis model, MSA-HA, achieved optimal values on the Acc_2, Acc_3, F1 score, and Corr metrics. On the MAE metric, it differed from the optimal value of 28.81% by only 0.1%. These results demonstrate the superiority of MSA-HA in the emotion detection task and confirm that it can meet the detection requirements for emotional driving. As indicated in Table 2, the prediction results of the multi-modal model surpass those of any single-modal model. Furthermore, in unimodal detection, the audio modality exhibited the worst performance, which suggests that the extraction of emotional features from the audio modality poses a significant challenge. This is a problem that needs to be addressed in future work.

Table 1 Comparison of performance between MSA-HA and mainstream multimodal sentiment analysis models

Model	ACC_2/%	ACC_3/%	F1/%	MAE/%	Corr/%
MLF_DNN	81.91	74.47	81.87	**28.81**	72.54
MLMF	80.56	73.50	80.67	29.73	72.03
SELF_MM	80.28	72.79	79.88	32.14	70.19
BERT-MAG	76.40	69.83	76.34	35.45	62.13
MMIM	76.21	68.47	75.98	35.89	59.80
TETFN	79.59	73.76	79.69	30.64	71.02
CENET	79.98	73.40	80.08	31.11	69.82
MSA-HA	**83.50**	**74.76**	**83.46**	28.91	**73.56**

Table 2 Comparison of single-mode and multi-mode detection performance

Model	ACC_2/%	ACC_3/%	F1/%	MAE/%	Corr/%
T	77.37	68.09	77.22	34.65	62.46
A	55.42	9.67	52.31	50.46	10.31
V	72.44	60.25	71.69	40.62	47.90
M	**83.50**	**74.76**	**83.46**	**28.91**	**73.56**

3.2.3 Ablation

To further validate the performance of the MSA-HA model, ablation studies were conducted with the training parameters set to a batch size of 128, an early stopping patience of 8, and a learning rate of 0.001. Initially, to verify the superiority of mLSTM in feature extraction, the mLSTM component in the MSA-HA was replaced with a standard LSTM, yielding the MSA-HA(LSTM) variant. As shown in Table 3, compared to MSA-HA(LSTM), the proposed MSA-HA achieved improvements of 1.00% and 2.42% in binary and ternary classification accuracies, respectively, along with increases of 1.01% in the F1-score and 0.72% in the correlation coefficient, and a decrease of 0.19% in the Mean Absolute Error (MAE), indicating that mLSTM possesses a stronger capability to capture long-range dependencies. Furthermore, to evaluate the effectiveness of the hybrid attention module, this module was removed to generate the MSA model, which directly utilizes text, audio, and visual features for tensor fusion. Comparative results in the Table3 indicate that the MSA-HA outperformed the MSA model with accuracy improvements of 3.52% (binary) and 1.36% (ternary), increases of 3.38% and 3.74% in the F1-score and correlation coefficient, respectively, and a 0.82% reduction in MAE; this demonstrates that the hybrid attention module enhances feature interaction across different modalities, thereby boosting overall performance. In summary, the integration of mLSTM for feature extraction and the hybrid attention module for feature interaction significantly improves the model's capability for multimodal sentiment recognition.

Table 3 Ablation results

Model	ACC_2/%	ACC_3/%	F1/%	MAE/%	Corr/%
MSA-HA (LSTM)	82.50	72.34	82.45	29.10	72.84
MSA	79.98	73.40	80.08	29.73	69.82
MSA-HA	**83.50**	**74.76**	**83.46**	**28.91**	**73.56**

3.3 Fatigue and distracted driving detection model

3.3.1 Evaluation indicators

For the driver fatigue and distraction detection task, the evaluation metrics adopted in this study are mean Average Precision (mAP), Parameter quantity and Flops.

In the YOLO series of object detection algorithms, mAP@0.5 and mAP@0.5-0.95 are two commonly used performance evaluation metrics for measuring the model's detection accuracy at different thresholds. These metrics are related to the model's prediction accuracy, recall rate, and various IoU (Intersection over Union) thresholds, playing a crucial role in evaluating the performance of object detection tasks. mAP@0.5 is the mean mAP calculated at a specific IoU threshold of 0.50. IoU is a metric that measures the degree of overlap between the predicted bounding box and the ground-truth box; a higher IoU value indicates a greater degree of overlap and thus a more accurate prediction. mAP@0.5-0.95 represents the mAP averaged over a range of IoU thresholds, from 0.5 to 0.95. It provides a more comprehensive evaluation metric by assessing the model's performance across multiple IoU thresholds, reflecting its capability under various overlap conditions (from lower to higher overlap). This metric offers a more holistic reflection of the detection model's accuracy and is more commonly used, especially in tasks that demand high precision.

3.3.2 Comparative experiments

To validate the effectiveness of the DPCLN, comparative experiments were conducted between the YOLOv8n-DPCLN and a series of other models, including YOLOv5s, YOLOv6n[25], YOLOv9s, YOLOv10n[26], YOLOv11n, YOLOv12n[27], and YOLOv8n. YOLOv8n was selected as the baseline model due to its superior comprehensive performance compared to other YOLO variants. The experimental results presented in Table 4 indicate that the YOLOv8n-DPCLN achieved the best performance, with mAP@0.5 and mAP@0.5-0.95 reaching 84.5% and 55.9%, respectively. Relative to the baseline YOLOv8n, the YOLOv8n-DPCLN improved mAP@0.5 and mAP@0.5-0.95 by 1.9% and 3.7%, respectively, while simultaneously reducing the parameter count by 0.4 M and the computational cost by 1.1 GFLOPs. Consequently, it can be concluded that the proposed DPCLN not only enhances the feature representation capability of the model but also improves computational efficiency and reduces the consumption of computational resources.

Table 4 Comparison of performance of different series models in fatigue and distracted driving detection

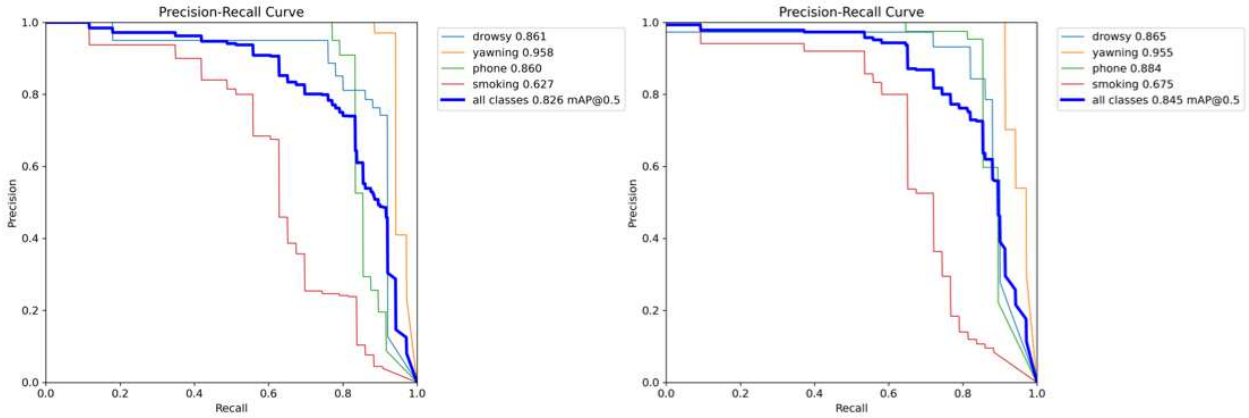
Model	mAP@0.5/%	mAP@0.5-90/%	Parameters/M	Flops/G
YOLOv5s	0.792	0.502	7.8	18.7
YOLOv6n	0.736	0.444	4.2	11.5
YOLOv9s	0.845	0.532	6.2	22.1
YOLOv10n	0.83	0.523	**2.3**	6.5
YOLOv11n	0.799	0.506	2.6	6.3
YOLOv12n	0.795	0.494	2.5	**5.8**

YOLOv8n(baseline)	0.826	0.522	3.0	8.2
YOLOv8n-DPCLN	**0.845**	**0.559**	2.6	7.1

Figure 6 compares the detection effects of the YOLOv8n and YOLOv8n-DPCLN models. It is evident that the detection performance of the YOLOv8n-DPCLN model is significantly superior to that of YOLOv8n. Figure 7 illustrates the Precision-Recall (PR) curves for both YOLOv8n and YOLOv8n-DPCLN, including the mAP@0.5 values for each class. Specifically, (a) represents the PR curve for the YOLOv8n model, while (b) represents that of the YOLOv8n-DPCLN model. With the exception of the "yawning" category, where the mAP@0.5 is slightly lower than that of YOLOv8n, the YOLOv8n-DPCLN model achieves higher mAP@0.5 values across all remaining categories. Based on the analysis of the detection performance and PR curves, it can be concluded that the YOLOv8n-DPCLN model satisfies the requirements for detecting driver fatigue and distraction.



Fig.6 Detection effect of YOLOv8n-DPCLN



(a) (b)

Fig.7 PR curve of YOLOv8n-DPCLN

4 System testing

4.1 Testing environment

The test environment for the MSDMS system was configured as follows: an NVIDIA RTX 2060S (8 GB) graphics card, an Intel Core i5-10400F @ 2.9 GHz CPU, 16 GB of ADATA RAM, and the Windows 10 operating system. The video capture device used was a Hikvision DS-E12.

4.2 Testing results

Figure 8 shows the detection performance of a simplified prototype of the MSDMS system. This system can perform real-time detection by connecting to a camera or by analyzing an uploaded video. The left side of the system interface serves as the image display area, while the right side features a "Connect Camera" button, an "Upload Video" button, and a "Detection Status" indicator. If an abnormal driving state is detected, it will trigger both text and audible alerts.Through extensive experimental testing of this system, the results have validated its effectiveness in handling emotional, fatigued, and distracted driving detection tasks. In the system testing experiments, five groups of videos were collected to evaluate the system's recognition rate for five class labels: "negative," "drowsy," "yawning," "smoking," and "phone." The "negative" label represents emotional driving. To validate the robustness of the MSA-HA model in cases of missing modalities, this category was further divided into three subgroups, representing scenarios where: (1) all three modalities are present, (2) only the visual modality is present, and (3) only the audio and text modalities are present. There were five video segments per subgroup. The "drowsy" and "yawning" labels represent fatigue driving, while "smoking" and "using phone" represent distracted driving.There were five video segments for each of these four labels, too. In each video segment, the actual number of occurrences for the state corresponding to each label was set to 20.

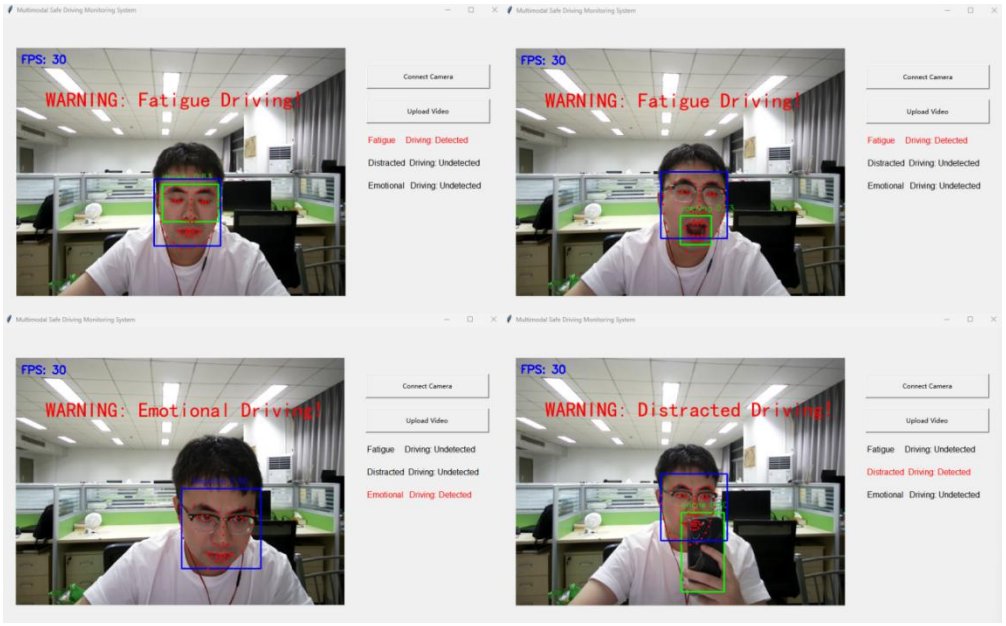


Fig.8 Detection effect of MSDMS

emotional driving test, the number of "negative" instances identified under different scenarios in the videos was counted to calculate the system's comprehensive recognition rate for emotional driving. Table 5 presents the detection results of the MSDMS system for emotional driving. It can be observed that in the scenario where all three modalities were present, the system's recognition rate(RR) was 84%,and the false alarm rate(FAR) is 2.3%. When only the visual modality was present, the recognition rate was 73%, and the false alarm rate is 5.2%. When only the text and audio modalities were present, the recognition rate was 79%.,and the false alarm Rate is 2.5%. This indicates that the model possesses a certain degree of robustness in scenarios with missing modalities. In summary, the system's comprehensive recognition rate for emotional driving is 78.7% ,and the false alarm rate is 3.3%.

Table 5 Emotional driving test results

Labels	Actual Counts	Recognition Counts	True Positive Counts	RR/%	FAR/%
Negative (**Trimodal**)	100	86	84	84	2.3
Negative (**Visual**)	100	77	73	73	5.2
Negative (**Visual+Audio**)	100	81	79	79	2.5
Overall	300	244	236	78.7	3.3

In the performance evaluation of the driver fatigue and distraction detection , the frequency of specific behaviors labeled as "drowsy," "yawning," "smoking," and "phone" are quantified from video recordings. The recognition rates for each corresponding state were calculated to determine the overall system efficacy for recognizing fatigued and distracted driving. Table 6 presents the detection results of the MSDMS. The system achieved recognition rates of 91% and 95% for "drowsy" and "yawning," respectively, yielding a composite recognition rate of 93% for fatigued driving. Furthermore, the recognition rates for "smoking" and "phone" were 80% and 92%, respectively, resulting in a combined recognition rate of 86% for distracted driving. These results validate the system's effectiveness in accurately identifying instances of driver fatigue and distraction.

Table 6 Detection results of fatigue and distraction driving

Labels	Actual Counts	Recognition Counts	True Positive Counts	RR/%	FAR/%
drowsy	100	93	91	91	2.2
yawning	100	95	95	95	0
smoking	100	80	80	80	0
phone	100	92	92	92	0

5 Conclusion

To provide a more comprehensive evaluation of driver behavioral status and further improve the accuracy and practicality of safety monitoring in construction machinery, this paper proposes a Multimodal Safe Driving Monitoring System (MSDMS). This system is designed to handle three detection tasks: emotional driving, fatigue driving, and distracted driving. Addressing the limitations regarding feature selection and the insufficient interaction capability among different features in current emotional driving detection, the Multimodal Sentiment Analysis Model based on Hybrid Attention (MSA-HA) is adopted for this task. Experimental results on the SIMSv2 dataset indicate that the MSA-HA model achieves an Acc-2, Acc-3, F1-score, and Precision Correlation (Corr) of 83.50%, 74.76%, 83.46%, and 73.56%, respectively, outperforming other mainstream multimodal sentiment analysis models. For the detection of fatigue and distracted driving, the YOLOv8n-DPCLN model is employed. Experimental results show that its mAP@0.5, mAP@0.5-0.95, parameter count, and computational cost are 84.5%, 55.9%, 2.6 M, and 7.1 GFLOPs, respectively. Compared to the baseline YOLOv8n, the YOLOv8n-DPCLN improves mAP@0.5 and mAP@0.5-0.95 by 1.9% and 3.7%, respectively, while reducing the parameter count by 0.4 M and computational cost by 1.1 GFLOPs. In the final system testing of the MSDMS, which integrates the aforementioned models, the recognition rates for emotional, fatigue, and distracted driving reached 78.7%, 93%, and 86%, respectively. These results indicate that the system is applicable for the comprehensive detection of abnormal driving behaviors. The MSDMS enhances the comprehensiveness of driving detection tasks and effectively improves detection efficiency, thereby ensuring the operational safety of construction machinery.

Acknowledgements. This work was supported by National Natural Science Foundation of China whose number is 62471241

Declarations

Funding

This work was supported by Natural Science Foundation of China[62471241].

Conflicts of interest

The authors have no relevant financial or non-financial interests to disclose.Availability of data and material

Ethics approval and consent to participate

This article does not contain any studies with human participants or animals performed by any of the authors

Consent for publication

Not applicable.

Availability of data and material

The SIMSv2 dataset analyzed during the current study is available in the [ch-sims-v2] repository, [https://github.com/search?q=ch-sims-v2&type=repositories]. The custom dataset generated during the current study is available from the corresponding author on reasonable request.

Code availability

The code used in this study is available from the corresponding author on reasonable request.

Authors' contributions

All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by Hao Cui, Jie Cai and Lingling Tian . The first draft of the manuscript was written by Hao Cui, Jie Cai, Lingling Tian, Zepeng Wei and Dengyin Zhang. All authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

References

[1] Bedi, J.K., Singh, R.P., Gupta, A.K.: Heavy machinery operators: necessary competencies to reduce construction accidents. IOP Conf. Ser. Earth Environ. Sci. 889(1), 012028 (2021). https://doi.org/10.1088/1755-1315/889/1/012028

[2] Hekmatmanesh, A., Dehghani, M., Handroos, A.: Biosignals in human factors research for heavy equipment operators: a review of available methods and their feasibility in laboratory and ambulatory studies. IEEE Access 9, 97466–97482 (2021). https://doi.org/10.1109/ACCESS.2021.3094829

[3] Abbas, Q., Ibrahim, M.E.A., Khan, S., Nasir, H.: Hypo-driver: a multiview driver fatigue and distraction level detection system. CMC-Comput. Mater. Contin. 71(1), 1999–2017 (2022). https://doi.org/10.32604/cmc.2022.020359

[4] Wang, Z.H., Chen, H.Y., Zhang, Z.H.: Intelligent monitoring system for safe driving based on convolutional neural network. Autom. Appl. 66(06), 36–39 (2025). https://doi.org/10.19769/j.zdhy.2025.06.012

[5] Dong, B.T., Lin, H.Y., Chang, C.C.: Driver fatigue and distracted driving detection using random forest and convolutional neural network. Appl. Sci. 12(17), 8674 (2022). https://doi.org/10.3390/app12178674

[6] Li, B., Sun, L.C.: A review of the impact of driving emotion on traffic safety and its recognition and regulation strategies. J. Saf. Environ. 24(10), 3909–3923 (2024). https://doi.org/10.13637/j.issn.1009-6094.2024.0401

[7] Ying, N., Jiang, Y., Guo, C., Zhang, Y.: A multimodal driver emotion recognition algorithm based on the audio and video signals in internet of vehicles platform. IEEE Internet Things J. 11(22), 35812–35824 (2024). https://doi.org/10.1109/JIOT.2024.3363176

[8] Sun, W., Liu, Y., Li, S., Zhang, J.: Research on driver’s anger recognition method based on multimodal data fusion. Traffic Inj. Prev. 25(3), 354–363 (2024). https://doi.org/10.1080/15389588.2023.2297658

[9] Baltrušaitis, T., Robinson, P., Morency, L.P.: OpenFace: An open source facial behavior analysis toolkit. In: 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 1–10. IEEE, Lake Placid (2016). https://doi.org/10.1109/WACV.2016.7477553

[10] Eyben, F., Wöllmer, M., Schuller, B.: Opensmile: the munich versatile and fast open-source audio feature extractor. In: Proceedings of the 18th ACM International Conference on Multimedia, pp. 1459–1462. ACM, Firenze (2010). https://doi.org/10.1145/1873951.1874246

[11] Eyben, F., Scherer, K.R., Schuller, B.W., Sundberg, J., André, E., Busso, C., Devillers, L.Y., Epps, J., Laukka, P., Narayanan, S.S., Truong, K.P.: The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. IEEE Trans. Affect. Comput. 7(2), 190–202 (2015). https://doi.org/10.1109/TAFFC.2015.2457417

[12] Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 4171–4186. Association for Computational Linguistics, Minneapolis (2019). https://doi.org/10.18653/v1/N19-1423

[13] Beck, M., Pöppel, K., Spanring, M., Schwenker, F.: xLSTM: Extended long short-term memory. arXiv preprint arXiv:2405.04517 (2024)

[14] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems, vol. 30. Curran Associates, Inc. (2017)

[15] Zadeh, A., Chen, M., Poria, S., Morency, L.P.: Tensor Fusion Network for Multimodal Sentiment Analysis. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pp. 1103–1114. Association for Computational Linguistics, Copenhagen (2017). https://doi.org/10.18653/v1/D17-1115

[16] Chollet, F.: Xception: Deep learning with depthwise separable convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1251–1258. IEEE, Honolulu (2017). https://doi.org/10.1109/CVPR.2017.195

[17] Liu, Y., Yuan, Z., Mao, H., Wu, C.: Make acoustic and visual cues matter: Ch-sims v2.0 dataset and av-mixup consistent module. In: Proceedings of the 2022 International Conference on Multimodal Interaction, pp. 247–258. ACM, Bangalore (2022). https://doi.org/10.1145/3536221.3556598

[18] Yu, W., Xu, H., Meng, F., Zhu, Y., Ma, Y., Wu, J., Zou, J.: Ch-sims: A Chinese multimodal sentiment analysis dataset with fine-grained annotation of modality. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 3718–3727. Association for Computational Linguistics, Online (2020). https://doi.org/10.18653/v1/2020.acl-main.340

[19] Liang, P.P., Zadeh, A., Casanueva, L.P.: Efficient low-rank multimodal fusion with modality-specific factors. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, pp. 2247–2257. Association for Computational Linguistics, Melbourne (2018). https://doi.org/10.18653/v1/P18-1212

[20] Yu, W., Xu, H., Yuan, Z., Wu, J.: Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, no. 12, pp. 10790–10797 (2021). https://doi.org/10.1609/aaai.v35i12.17282

[21] Rahman, W., Hasan, M.K., Lee, S., Zadeh, A., Mao, C., Morency, L.P., Hoque, E.: Integrating multimodal information in large pretrained transformers. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pp. 2359–2369. Association for Computational Linguistics, Online (2020). https://doi.org/10.18653/v1/2020.emnlp-main.187

[22] Han, W., Chen, H., Poria, S.: Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pp. 8475–8488. Association for Computational Linguistics, Online (2021). https://doi.org/10.18653/v1/2021.emnlp-main.723

[23] Wang, D., Guo, X., Tian, Y.: TETFN: A text enhanced transformer fusion network for multimodal sentiment analysis. Pattern Recognit. 136, 109259 (2023). https://doi.org/10.1016/j.patcog.2022.109259

[24] Wang, D., Liu, S., Wang, Q., Tao, D.: Cross-modal enhancement network for multimodal sentiment analysis. IEEE Trans. Multimed. 25, 4909–4921 (2022). https://doi.org/10.1109/TMM.2022.3184236

[25] Li, C., Li, L., Jiang, H., Weng, K., Geng, Y., Li, L., Ke, Z., Li, Q., Cheng, M., Nie, W., Li, Y., Zhang, B., Lu, Y., Yang, R., Xu, C., Chu, X., Wei, X., Wang, X.: YOLOv6: A single-stage object detection framework for industrial applications. arXiv preprint arXiv:2209.02976 (2022)

[26] Wang, A., Wang, L., Wang, L., Hao, Y.: YOLOv10: Real-time end-to-end object detection. arXiv preprint arXiv:2405.14458 (2024)

[27] Tian, Y.J., Ye, Q.X., Doermann, D.: YOLOv12: Attention-centric real-time object detectors. arXiv preprint arXiv:2502.12524 (2025)