ISSN 1330-3651 (Print), ISSN 1848-6339 (Online)
Received: 11 July 2025; Accepted: 29 February 2025

https://doi.org/10.17559/TV-20250620002763
Original scientific paper

Intelligent Detection of Road Rage Using PNN Parameter Optimization and Multimodal Driver Speech and Text Data

EnLin XIE*, YiLiu HUANG

Abstract: Road rage is a critical factor in traffic accidents, often expressed through vocal and semantic cues. This study proposes a multimodal road rage detection system that integrates speech and text features. Speech signals are processed by a probabilistic neural network (PNN) optimized with an Improved Sand Cat Swarm Optimization (ISCSO) algorithm, while text features are modelled with a long short-term memory (LSTM) network. A decision-level fusion strategy combines outputs from both modalities. Experiments on a self-constructed dataset of 10,000 speech samples and corresponding text corpora demonstrate that the proposed model achieves superior performance compared to CNN, DBN, TextCNN, and hybrid deep learning baselines. The system achieved maximum accuracy and recall of 95.61% and 99.31%, while maintaining low computational overhead (minimum detection time 58 ms, memory usage 10.06%). These findings suggest that the ISCSO-PNN and LSTM multimodal fusion framework provides an efficient and effective approach to detecting road rage, with strong potential for integration into real-time intelligent transportation systems.

Keywords: ISCSO; parameter optimization; PNN; road rage detection; speech; text

1 INTRODUCTION

Road rage refers to the anger and inappropriate dangerous behavior exhibited by drivers in specific situations. Road rage behavior is an increasingly prominent problem in modern transportation, especially in recent years with the continuous increase in the number of motor vehicles, traffic congestion, and driving pressure, leading to more frequent extreme emotional reactions among drivers [1]. Related studies have also shown that road rage behavior not only triggers dangerous driving behaviors such as speeding and forced lane changes, but also exacerbates conflicts between drivers, posing a serious threat to public safety [2]. At present, with the development of intelligent transportation systems, emotion recognition such as speech and vision has become a hot topic in traffic safety research [3]. Compared to visual detection, speech recognition is more adaptable to complex environments, less susceptible to environmental influences, and has better recognition accuracy [4]. At present, multi-dimensional data recognition and deep learning have wide applications in the field of vehicle safety driving. However, traditional video recognition has low accuracy and problems such as misidentification. Therefore, to improve the effectiveness of road rage recognition, a research proposes an intelligent road rage detection technology based on speech data, which can warn drivers of dangerous behaviors in advance and reduce the risk of traffic accidents. This study has two innovations. Firstly, it starts with bimodal data of speech and text, comprehensively extracts the emotional characteristics of drivers, and provides a richer information foundation for road rage detection. The second is to study the use of Improved Sand Cat Swarm Optimization (ISCSO) to optimize the Probabilistic Neural Network (PNN) detection model and LSTM network for speech and text data fusion processing, to improve the performance of road rage detection. This study will provide technical support for safe driving of vehicles.

2 RELATED WORK

At present, the increasing frequency of road rage behavior has caused severe adjustments to road traffic safety. How to effectively detect road rage has become a focus of current social research. Related scholars have conducted extensive research on driver emotion recognition, such as Arumugam S conducted research on driver hazard detection and proposed a machine learning based multimodal automatic road anger detection technology. This technology combines GPS positioning and wearable devices to capture changes in the user's heart rate. By monitoring and categorizing drivers' abnormal heart rates, their risky behaviors can be determined. Tests have shown that this technology has good application effects, but for complex scenarios, the effect is generally moderate [5]. Hmidi N. et al. conducted research on emotion estimation of masked faces. It developed a two-part system: first, using CNN to create a system for identifying sanitary masks; secondly, develop an emotion estimation system that estimates the emotions of masked faces through three steps: facial element detection, feature point localization, and classification. In the study, Viola and Jones algorithms were used for facial element detection, and techniques such as SVM, KNN, and deep learning were employed for emotion estimation. By comparing the results of different methods and paying special attention to the impact of masks on performance, valuable insights and improvement directions have been provided for emotion recognition during the pandemic [6]. Liu P. et al. conducted research on the risk issues of modern mixed traffic and proposed a road rage detection technology based on mental models and machine learning to effectively reduce traffic risks, aiming to monitor pedestrian risks on the road. This technology supervises and recognizes the psychological state and language of drivers, and judges the risks of drivers through language classification models and emotional fluctuations. Tests have shown that this technology can accurately judge the driver's emotional judgment and effectively monitor the driver's road rage status [7].

In recent years, intelligent road rage detection technology based on deep learning has gradually emerged. The powerful feature extraction ability of deep learning provides a new approach for accurately identifying road rage emotions in driver speech. Yan L. studied the problem

of identifying environmentally friendly driving behaviors with the aim of reducing driving risks. A driving simulation platform was constructed and experiments were conducted. Based on significant driving behaviors related to fuel consumption, the segmented linear representation (PLR) method was used to fit multivariate time series. Extract time series features and input them into a random forest (RF) model to identify ecological driving behavior. The results indicate that the depth of the accelerator pedal, clutch pedal, brake pedal, steering wheel angle, and gear have a significant impact on fuel consumption. The prediction accuracy of the ecological driving behavior recognition model is superior to similar technologies, and an ecological driving behavior map during lane changing has been established, providing theoretical support for the development of ecological driving intervention measures [8]. Zhang Y. studied a new algorithm for identifying vehicles in an environment using 3D LiDAR point cloud data. A point cloud compression method is proposed to address the shortcomings of duplicate points, redundant points, and unordered point clusters in point cloud data, combined with nearest neighbor point and octree voxel center point boundary extraction techniques. Afterwards, a vehicle point cloud recognition algorithm based on image maps was used for vehicle recognition. Through testing on the KITTI dataset, the results show that the accuracy of this algorithm is higher than other methods, providing a more efficient solution for vehicle recognition in autonomous driving and 3D reconstruction, and improving the performance and recognition speed of hardware systems in processing point cloud data [9]. Hasan M. A. et al. studied the problem of facial emotion recognition and proposed a structural model based on YOLO face detection and feature extraction for predicting facial emotions. This model classifies facial images into seven emotions (natural, happy, sad, angry, surprised, fearful, disgusted). The experiment was based on the FER2013 dataset, and the results showed that the system accuracy reached 94%, verifying the robustness and speed of the model. This study provides effective methods for the application of emotion recognition in fields such as human-computer interaction, safety, and health [10]. Surana A. et al. conducted research on anger emotion detection in audio emotion recognition (AER). They used audio files from the CREMA-D dataset and proposed a hybrid algorithm of artificial neural network (ANN) and fuzzy logic, combined with specialized preprocessing techniques. The experimental results show that the algorithm outperforms traditional neural network methods in anger emotion detection, with higher efficiency and accuracy. The study also discussed the limitations and improvement suggestions of the experimental setup, emphasizing its potential application value in the fields of public safety and mental health [11].

According to the above research, it can be found that road rage driving behavior has a serious impact on traffic safety, and even leads to major traffic accidents. Therefore, in order to effectively detect driver's road rage, a research based on voice data and text data proposes an intelligent detection technology for road rage based on PNN parameter optimization and multimodal driver's voice and text data, in order to ensure the safety and effectiveness of road traffic.

3 CONSTRUCTION OF A ROAD RAGE DETECTION MODEL CONSIDERING BOTH SPEECH AND TEXT MODALITIES

3.1 Speech Feature Data Processing and Analysis

In the field of traffic safety, road rage drivers are in an irrational driving state and are prone to dangerous driving problems due to emotional outbursts during road driving. Related studies have shown that 90% of road rage drivers will express their emotions through language [12]. Therefore, a research proposes an intelligent road rage detection technology that integrates speech and text. This technology considers both voice and text information separately, and achieves accurate anger detection by fusing the two types of information. The technical process is shown in Fig. 1.

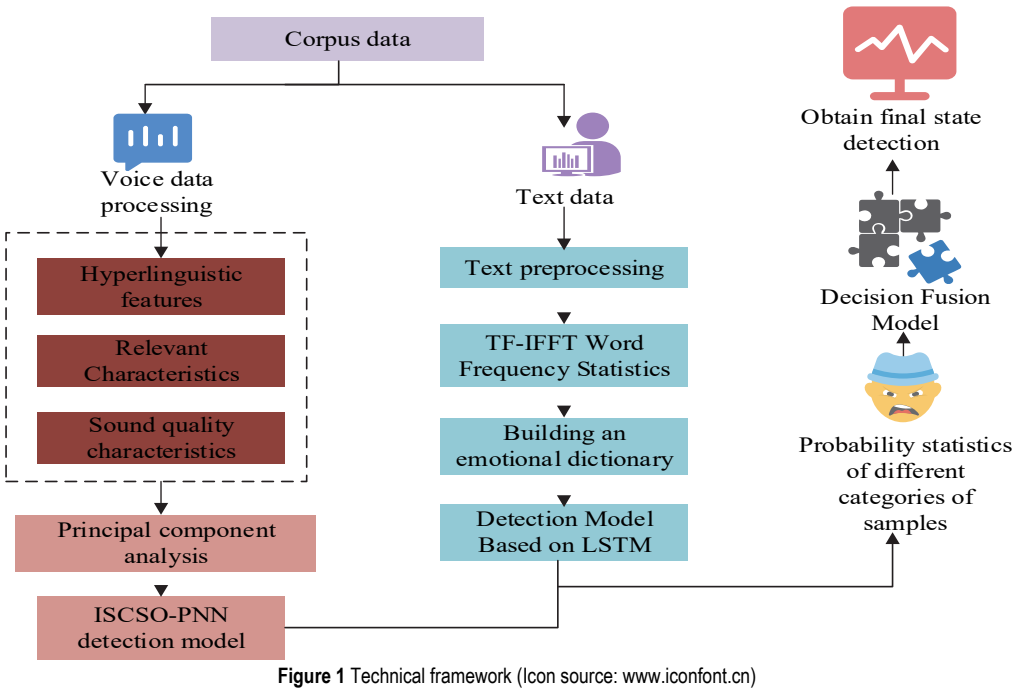


Figure 1 Technical framework (Icon source: www.iconfont.cn)

In Fig. 1, the technology utilizes two types of multimodal data, speech and text, as inputs. Subsequently, the two multimodal models are used for road rage state recognition, and finally, intelligent detection of driver road rage is achieved through decision-making. However, the study did not adopt feature level fusion, mainly because there is a large difference in the dimensionality of speech and text features. Speech contains acoustic features such as Mel coefficients, while text contains semantic features such as word vectors. Direct feature level fusion can easily lead to feature redundancy and noise accumulation. Therefore, decision level fusion is adopted to process two types of data separately and output decision results, retaining the advantages of single modality, avoiding cross modal feature mismatch problems, and improving detection performance. Therefore, the next step will be to process and analyze the speech data. In speech signal preprocessing, sensors detect that the raw speech data is affected by environmental noise and device interference. Therefore, preprocessing is performed next to remove noise. The study uses pre emphasis processing to compensate for high-frequency component attenuation, suppress low-frequency noise, and improve speech clarity. The formula is shown in Eq. (1) [13].

$$H(n) = 1 - \mu n^{-1} \quad (0.9 \le \mu \le 1) \tag{1}$$

In Eq. (1), n represents the index of the discrete speech signal sequence; μ represents the pre emphasis coefficient, which controls the high-frequency gain intensity and defaults to 0.95. Through pre emphasis processing, it can better represent the rapid and sharp speech in road rage state, with significantly increased high-frequency energy. Pre emphasis can enhance such features, as shown in Fig. 2 [14].

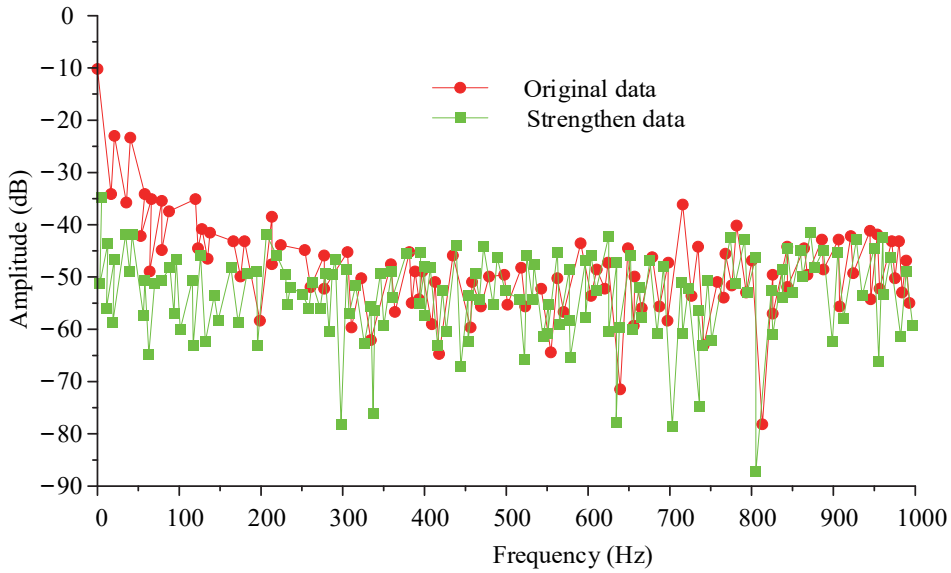


Figure 2 Pre emphasis processing effect of speech signal

In Fig. 2, there is significant signal enhancement in the rapid speech parts to ensure better discrimination of the driver's emotional state in that environmental state. Then, frame segmentation and windowing are used to segment non-stationary speech into short-term stationary frames to reduce spectral leakage, as shown in Eq. (2).

$$W(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right) \quad (0 \le n \le N-1) \tag{2}$$

In Eq. (2), $W(n)$ represents the window function. N represents the frame length, where the frame rate corresponds to a time period of 20-30 ms and a sampling size of 160-240 points @ 8 kHz. n represents the index of sampling points within the window function. Edge mutations are smoothed using split frames with windows to ensure speech continuity and avoid information loss [15]. In addition, endpoint detection is used to locate the effective start and end points of speech, and silence and noise segments are removed. This process is expressed using a short-term autocorrelation function, as shown in Eq. (3) [16].

$$R_w(k) = \sum_{m=0}^{K-1-k} S_w(m) \cdot S_w(m+k) \tag{3}$$

In Eq. (3), $S_w(m)$ represents the m-th sampling point after windowing; $R_w(k)$ is a short-term autocorrelation function; k represents the number of delay points, with a range of $0 \le k \le K_{\max}$; K represents the maximum number of delay points. By utilizing endpoint detection, it is possible to accurately capture roaring segments. Next, speech feature analysis will be conducted, focusing on three feature surfaces: hyper linguistic features, spectral correlation features, and sound quality features, to achieve recognition of the driver's road rage state [17]. In hyper linguistic feature analysis, short-term energy is used to reflect, as shown in Eq. (4).

$$E(i) = \sum_{n=0}^{N-1} S_i(n)^2 \tag{4}$$

In Eq. (4), $S_i(n)$ represents the i-th frame signal and N represents the frame length. Using N to reflect speech intensity, energy significantly increases in road rage state.

Next, the fundamental frequency F_0 is used to represent the frequency of vocal fold vibration, and the fundamental frequency jitter increases during road rage, as shown in Eq. (5) [18].

$$F_0 = \frac{f_s}{T_0} \tag{5}$$

In Eq. (5), T_0 represents the fundamental period obtained by cepstral method; T_0 represents the sampling rate (8 kHz). In spectral correlation feature analysis, the main purpose is to simulate the auditory characteristics of the human ear. During road rage, high-frequency energy is extended to 8 kHz, and Mel-Frequency Cepstral Coefficients (MFCC) are used to represent it, as shown in Eq. (6) [19].

$$\mathrm{MFCC}(q) = \sum_{m=1}^{M} \log[E(m)] \cdot \cos\left[\frac{\pi q(2m-1)}{2M}\right] \tag{6}$$

In Eq. (6), $E(m)$ represents the energy of the m-th Mel filter; q represents the cepstral coefficient number; M represents the number of Mel filters. The analysis of sound quality characteristics mainly considers resonance peaks, which can characterize the structure of the vocal tract, and the displacement of resonance peaks increases during road rage. Finally, the principal component analysis method is used to reduce the dimensionality and normalize the feature data. The feature matrix is normalized to $Z = A^T A$, and A is the normalized feature matrix. The above research has completed the processing and analysis of speech data.

3.2 Text Feature Data Processing and Analysis

After completing the processing and analysis of speech data, as machine learning models cannot directly use speech data to identify human emotions, further processing and analysis of speech text data is required. The specific technical process is shown in Fig. 3.

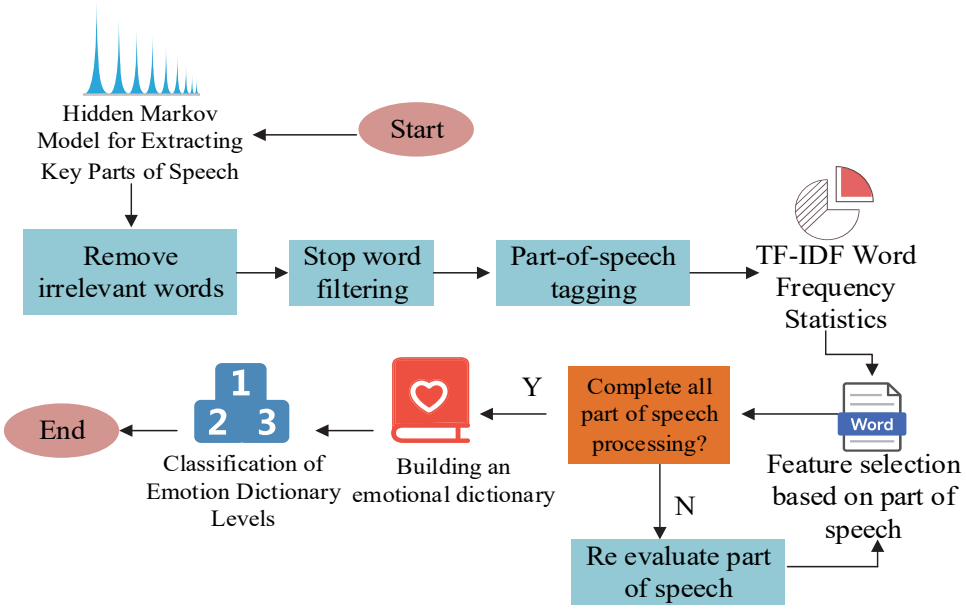


Figure 3 Process and analysis flow of text data

In the technical process of Fig. 3, the research performs basic preprocessing on the driver's expected text data, including deleting stop words and word segments that affect recognition, and annotating keyword properties. Next, the study uses the Term Frequency-Inverse Document Frequency (TF-IDF) algorithm to statistically analyze the frequency of different parts of speech texts, and further processes the text data using a sentiment dictionary to obtain text feature data that better reflects the emotional state of drivers. The text features reveal aggressive expressions at the semantic level, mainly based on 8000 road rage/non-road rage driving corpora. Firstly, text data preprocessing is carried out to solve the problem of Chinese no segmentation, extract parts of speech units with emotional value, and study the use of hidden Markov models for representation, as shown in Eq. (7) [20].

$$P(w_1, w_2, \ldots, w_m) \approx \prod_{i=1}^{m} P(w_i w_{i-1}) \cdot P(\mathrm{tag}_i w_i) \tag{7}$$

In Eq. (7), w_i represents the w_i-th word; tag_i represents the corresponding part of speech tag. In segmentation, continuous sentences are broken down into independent semantic units to solve the inherent problem of Chinese without space separation. Then dwell word filtering is performed to eliminate redundant words with no emotional value, as in Eq. (8) [21].

$$\mathrm{FilteredText} = \{w_i \mid w_i \notin \mathrm{StopSet}, i = 1, 2, \ldots, N\} \tag{8}$$

In Eq. (8), StopSet represents a predefined stop word list, such as "oh", "um", "again", etc. Part of speech tagging is based on Conditional Random Field (CRF), which identifies grammatical categories to support sentiment weighting as shown in Eq. (9).

$$P(y_r \mid x_r) = \frac{1}{Z(x_r)} \exp\left(\sum_k \lambda_k f_k(y_{i-1}, y_i, x, i)\right) \tag{9}$$

In Eq. (9), x_r represents the input word sequence; y_r represents the output part of speech tag sequence; f_k represents the characteristic function. Based on the analysis of word features, to better enhance the ability to represent road rage states, an emotional dictionary is constructed. The study uses TF-IDF word frequency statistics to evaluate the category discrimination of words as shown in Eq. (10) [22].

$$\mathrm{TF}_{\alpha,\beta} = \frac{f_{\alpha,\beta}}{\sum_{\chi=1}^{I} f_{\chi,\alpha}} \tag{10}$$

In Eq. (10), $f_{\alpha,\beta}$ represents the number of times the word β appears in sentence α, and the relative frequency of the word in a single sentence is calculated to reflect its local importance. Next, the inverse document frequency is calculated, as shown in Eq. (11).

$$\mathrm{IDF}_{\beta} = \log \frac{G}{G_{\alpha\beta}+1} \tag{11}$$

In Eq. (11), G represents the total number of sentences; $G_{\alpha\beta}$ represents the number of sentences containing the word β. Next, a sentiment dictionary is generated using part of speech based feature filtering. The specific steps are to calculate the TF-IDF mean by grouping words according to their parts of speech, including nouns, verbs, adjectives, etc. From each vocabulary, high-frequency words are selected to construct the sentiment dictionary, as shown in Tab. 1.

Table 1 Emotional vocabulary list

Part of speech category	Example of road rage state words	Non road rage state words	Negative emotional vocabulary intensity
Nouns	Silly, garbage, scum, madman, stuffing dog, porcelain bumpkin	Weather, roads, passenger seat, journey, green trees, service area	6~9
Verbs	Death by collision, pushing and shoving, stealing lanes, forcing to stop, stuffing, reincarnation, scratching, cursing	Driving, traveling, yielding, waiting, waiting, slowing down, starting	7~9
Adjective	Corrupt, shameless, impatient, disgusting, foolish, incompetent, barbaric	Smooth, safe, comfortable, pleasant, spacious, bright, peaceful	5~8
Special Words	His mother, grass mud horse, go to hell, fuck off, idiot	Drive carefully, have a smooth journey, pay attention to safety, thank you for giving way, please proceed first, don't worry	8~9

Tab. 1 covers four types of vocabulary with different parts of speech, ranging from level 1 to level 10. The higher the value, the higher the level of provocation. By processing and analyzing the high-frequency vocabulary mentioned above, the emotional state of the driver can be well identified.

3.3 Construction of Road Rage Intelligent Detection Model Based on Multimodal Data

After completing the processing and analysis of text and language datasets, a road rage intelligent detection technology combining text and speech is proposed. This process combines speech and text data to obtain the final intelligent road rage state detection result through fusion decision-making. The speech detection technology process is shown in Fig. 4.

In Fig. 4, the study utilizes ISCSO to optimize PNN parameters in speech data and constructs a road rage detection model based on ISCSO-PNN. The PNN optimized by ISCSO is sensitive to speech features, computationally efficient, and can quickly capture short-term high-frequency features of road rage speech. Among them, the study uses PNN probability density calculation to measure the similarity between samples and class centers, with the core being probability density estimation, as shown in Eq. (12).

$$\phi_{aj}(x_b) = \frac{1}{(2\pi)^{b/2}\sigma^b A} \exp\left[-\frac{(x-x_{aj})^T (x-x_{aj})}{2\sigma^2}\right] \tag{12}$$

In Eq. (12), x_b represents the input feature vector (31 dimensions); x_{aj} represents the j-th sample center of class a; σ represents the smoothing factor; A represents the number of samples within the class. The study calculates the similarity between samples and class centers through PNN, outputs the probability density, and uses the result with the largest output layer as the final result for detecting the anger state in the speech data. However, in PNN probability density calculation, it is easily affected by the size of the smoothing factor σ, such as σ being too small, which leads to overfitting and noise sensitivity in classification; Excessive σ can lead to blurred inter class boundaries and affect detection accuracy. To solve this problem, the ISCSO algorithm is used for parameter optimization. The parameter problem of SCSO algorithm is improved by using dynamic sensing radius and Levy flight strategy. The sensing radius update and position update of SCSO are shown in Eq. (13) [23].

$$\begin{cases} r_G = \alpha \cdot |\, Best_t - Worst_t \,| \\ X_{\mathrm{new}} = X_t + r_G \cdot Levy(\beta) \cdot (Best_t - X_t) \end{cases} \tag{13}$$

In Eq. (13), r_G represents the perceived radius of the sand cat; α is the contraction factor; $Levy(\beta)$ represents the random step size of the Levy distribution to avoid local optima; X_{new} represents the update location; $Best_t$ represents the current optimal solution position. In addition, the study uses minimizing the root mean square error of PNN to ensure the minimum error, as shown in Eq. (15) [14].

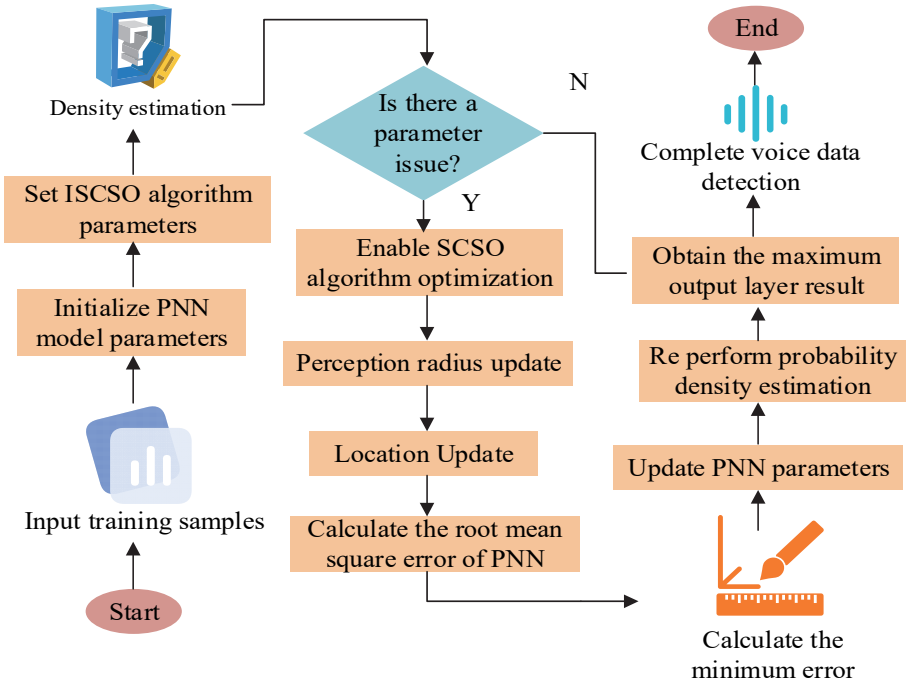


Figure 4 Process of speech detection technology

$$\mathrm{RMSE}=\sqrt{\frac{1}{N}\sum_{a=1}^{N}(y_a-o_a)^2} \tag{14}$$

In Eq. (14), y_a represents the predicted output, o_a represents the true label, and N represents the number of samples. Next, in text data detection, the LSTM gating mechanism was used to capture the long-term dependencies of text sequences, without using Transformer and attention mechanism models. Although both models are longer than long sequence modeling, they have high computational complexity and insufficient adaptability to real-time driving scenarios. LSTM excels at handling temporal dependencies of text sequences, effectively mining contextual associations of text emotions, and better adapting to the low latency and high response requirements of driving scenarios, thus meeting the requirements of object detection. It uses a forget gate to filter redundant information, as shown in Eq. (15) [25].

$$A_t=\sigma(Q_A\cdot[h_{t-1},x_t]+g_A) \tag{15}$$

In Eq. (15), σ is the activation function, h_{t-1} is the previous hidden layer state, g_A is bias vector, and x_t is the current input. The hidden layer is shown in Eq. (16).

$$h_t=u_t\odot\tanh(C_t) \tag{16}$$

In Eq. (16), x_t represents the forget gate output; u_t represents the weight of the output gate; C_t represents the state of the memory unit; $\odot$ stands for Hadamard product. Finally, the study utilizes the adaptive weight method to integrate two types of decisions and achieve intelligent detection of road rage. In order to overcome the limitations of fixed weights, category probability adaptive weights were designed, which dynamically allocate weights and use high confidence mode as the dominant decision. The fusion decision is shown in Eq. (17) [26].

$$\mathrm{result}=\arg\max_j\left(\mu_j\cdot\left(p_j^W+p_j^V\right)\right) \tag{17}$$

In Eq. (17), argmax_j is the decision function. In detection, if the confidence level of the speech modality for "road rage" reaches 0.9, the speech weight will be increased and used as the final detection result.

4 PERFORMANCE VERIFICATION OF UNIMODAL AND MULTIMODAL ROAD RAGE DETECTION MODELS

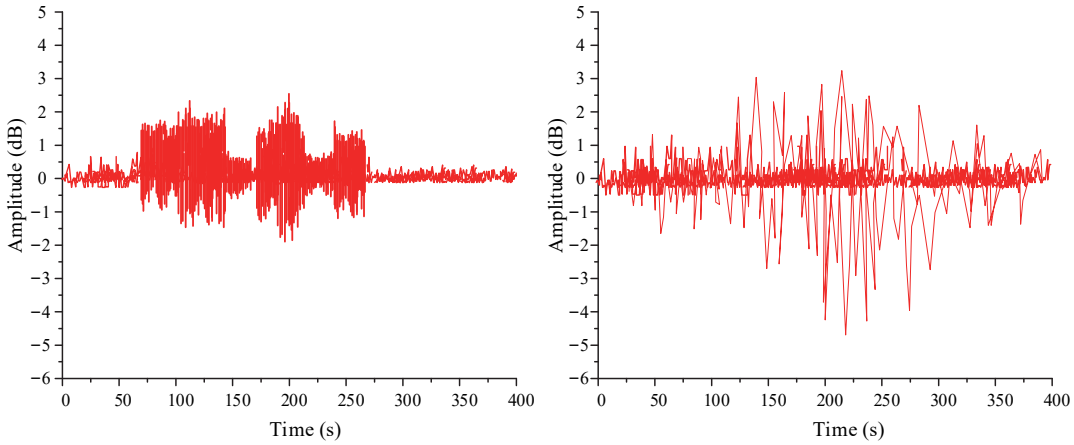
4.1 Performance Verification of Unimodal ISCSO-PNN and LSTM Models

Next, to test the technology proposed by the research, corresponding experiments were conducted. The experimental environment configuration is shown in Tab. 2.

Table 2 Experimental environment settings

Environmental parameters	Configuration
Processor	Intel Xeon E5-2690 v4
Memory	56 GB DDR4
Graphics card	NVIDIA Tesla P100
Hard disk	2 TB
operating system	Ubuntu 20.04 LTS
programming language	Python 3.8
Deep learning framework	TensorFlow 2.4.1 and PyTorch 1.8.1

Tab. 2 shows the experimental environment parameters, including the settings of software and hardware environment parameters. In addition, experimental data preprocessing and feature extraction tools use NumPy, SciPy, and Librosa libraries. In addition, Matplotlib and Seaborn libraries were used for system visualization in the experiment. In the experiment, a total of 10000 speech data were recorded through real self driving scenario simulation, of which 40% were road rage speech samples. This dataset covers different genders (62% male, 38% female) and age groups (45% aged 20-30, 40% aged 31-50, and 15% aged 51 and above), including Mandarin and 8 dialect variants, such as Sichuan dialect, Cantonese, and Shaanbei dialect variants. The data comes from 60% simulated driving scenarios and 40% real road recordings, which can better reflect the voice characteristics of road rage among different groups. In addition, text transcription data is preprocessed using the Librosa library and combined with the Transformer model for automatic speech recognition and transcription, converting speech into text. The final text data is then segmented and proofread using Python's NLTK tool. All speech data were in WAV mono format. The training process dataset was divided into training set, validation set, and testing set in a ratio of 7:2:1 to ensure the adequacy and generalization ability of the model training. Next, the study tested the effectiveness of the ISCSO-PNN model in speech and text, as well as the LSTM model combined with TF-IDF, in a unimodal scenario. Among them, Convolutional Neural Network (CNN), Deep Belief Network (DBN), and Text Convolutional Neural Network (TextCNN) were introduced as testing benchmarks. In the speech scene, the study selected simple and complex sound data for testing, as shown in Fig. 5.



(a) Simple scene sound amplitude (b) Complex scene sound amplitude

Figure 5 Features of audio data in simple and complex scenarios

Fig. 5a shows the amplitude of a simple scene audio, with a relatively concentrated overall fluctuation and a small amplitude range. In the complex scene sound amplitude shown in Fig. 5b, there was a significant amplitude fluctuation between 100 s and 350 s, indicating that the scene recorded audio data of the driver's road rage state. In the driver's road rage scene shown in Fig. 5b, it can be observed that there are significant fluctuations in the audio within 400 seconds of monitoring, and some fluctuations are significantly higher than those in the conventional scene. This indicates that the driver's road rage state voice information fluctuates abnormally, making it prone to behaviors such as honking, insults, and dangerous driving, increasing road safety risks. Therefore, the study selected two scenarios for experimentation and added Gaussian noise in complex scenes to verify the effectiveness of different techniques, as shown in Fig. 6.

In Fig. 6, three threshold ranges were set for categories 0, 1, and 2, where 0 represents unrecognized, 1 represents non road rage state, and 2 represents road rage state. In the simple scenario shown in Fig. 6a, the CNN experienced one instance of unrecognized and four instances of misidentification as road rage. The DBN model did not encounter any unrecognized situations during detection, but there were three instances of misidentification. In contrast, FA-PNN did not exhibit any unrecognized issues and was consistent with the true values. It can be observed that the FA-PNN model can accurately determine the road rage state of drivers in simple scenarios, covering 80% of road rage situations. Once the driver is diagnosed with road rage, the vehicle system will be able to provide effective warning feedback, avoiding the driver from engaging in more extreme behavior and affecting road traffic safety. Fig. 6b shows the results of a complex noisy scene, where CNN experienced 4 instances of recognition and 6 instances of misidentification. The DBN model also experienced 6 instances of misidentification during the recognition process, but there were no instances of misidentification. The FA-PNN with the best overall performance only had one misidentification at 36 s, indicating the best overall performance. The FA-PNN model optimizes the PNN parameters to avoid overfitting and inter class boundary blurring, enhancing the capture of high-frequency features of road rage speech. However, CNN and DBN have not been optimized for speech noise, which interferes with feature extraction and leads to higher recognition errors. In complex scenarios, such as vehicles honking or drivers making noise, environmental monitoring data is difficult to determine. However, research technology can accurately extract target sound sources and filter out unnecessary noise, allowing abnormal driver behavior to be monitored in advance and reducing road risks. Next, the study selected four types of part of speech data from the sentiment dictionary to test the recognition accuracy of different models in text scenes, as shown in Fig. 7.

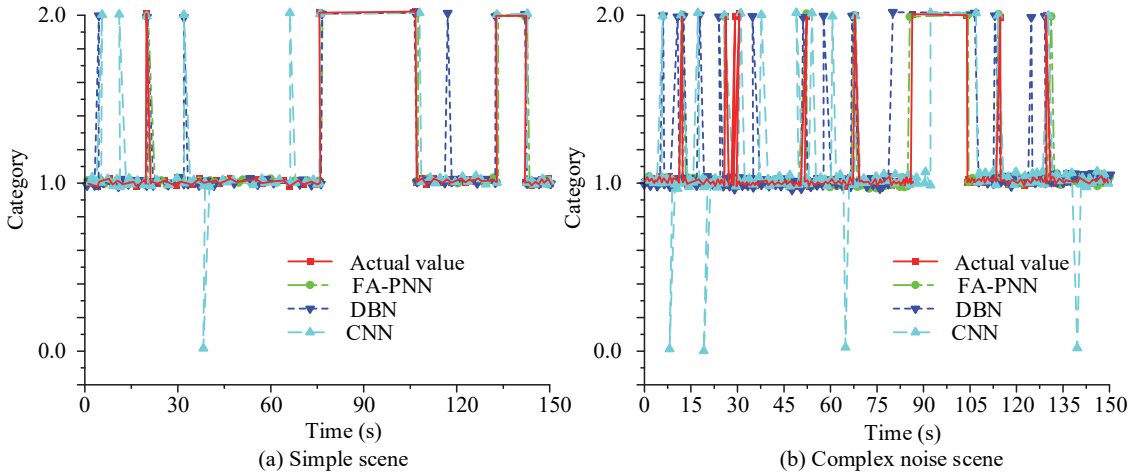


Figure 6 Speech road rage state discrimination test in different scenarios

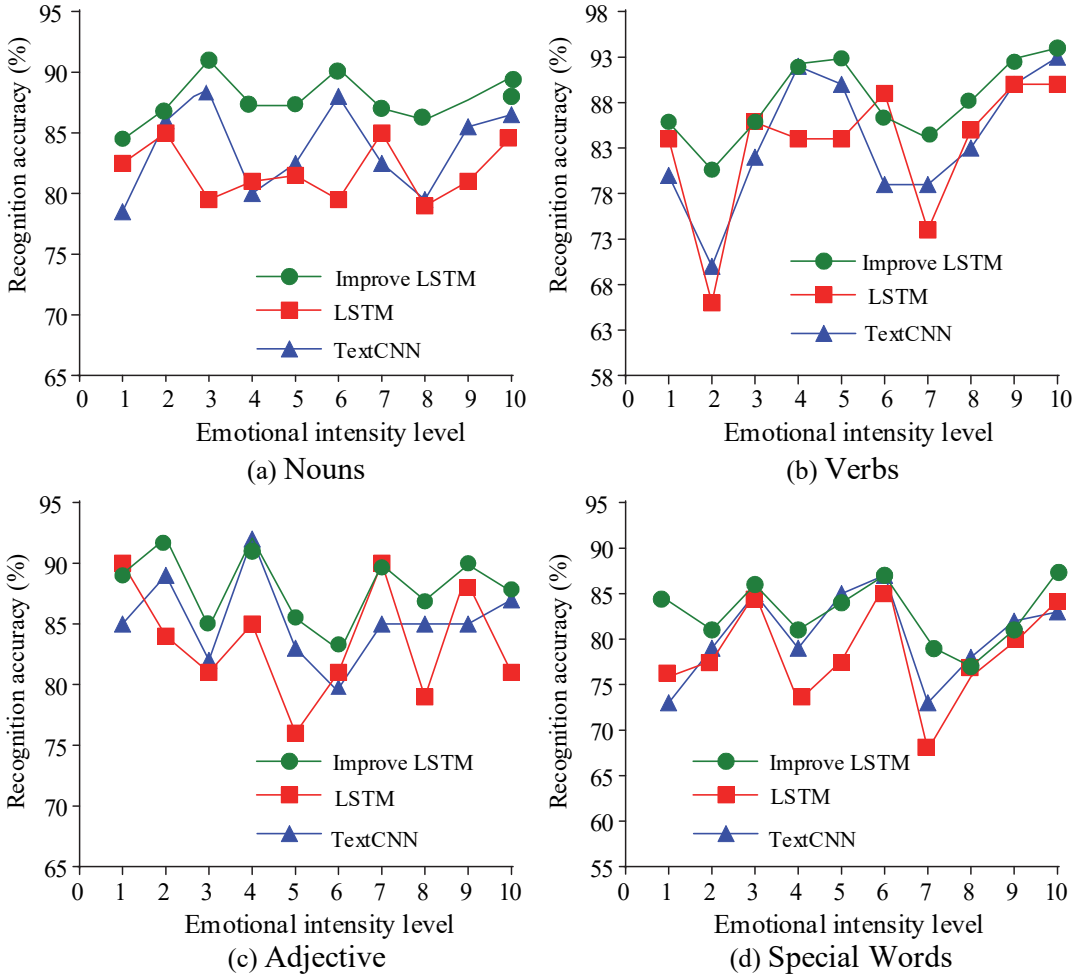


Figure 7 Text data road rage state recognition test

Fig. 7a shows the test results of noun data, with 10 emotion levels set. In this scenario, the improved LSTM had the highest accuracy with an average of 88.6%, followed by LSTM with an average recognition accuracy of 83.2%, and TextCNN performed better with an average recognition accuracy of 82.8%. In verb data, both LSTM and TextCNN exhibited significant errors in level 2 sentiment recognition, with an overall recognition accuracy of less than 70%. The improved LSTM performed the best overall at 81.0%. In the summary of adjective data testing, LSTM showed significant fluctuations and performed the worst overall, with an average recognition accuracy of 80.8%, while TextCNN had an average recognition accuracy of 85.7%. The overall best performing improved LSTM had an average recognition accuracy of 87.6%. Finally, in the special vocabulary data in Fig. 7d, the overall recognition performance of all three models showed a significant decrease. The best performing model was the improved LSTM, with an average recognition accuracy of 84.2%, while LSTM and TextCNN were 75.3% and 78.7%, respectively. Improving LSTM by effectively capturing the long-term dependencies of noun sequences through gating mechanisms, combined with the clear emotional

tendencies of nouns in the sentiment dictionary, such as aggressive words like 'idiot', enhances recognition accuracy. However, special vocabulary is often colloquial and unstructured expressions such as "grass mud horse", which other models find difficult to capture in their contextual associations, resulting in decreased performance. It can be observed that the proposed technology can make abnormal judgments on most dangerous and emotional words, especially when drivers lose control of their emotions, which is beneficial for timely intervention in road risk behaviors such as verbal abuse, provocation, and dangerous driving.

4.2 Performance Verification of Dual-Mode Detection Model Combining ISCSO-PNN and LSTM

After completing the testing of unimodal scenarios, the study tested the performance of different models in multimodal scenarios (speech + text). The study adopted an improved LSTM combination model scheme combining ISCSO-PNNJ, while introducing DBN-LSTM and CNN-LSTM combination schemes for comparison. In the road rage state data that integrated speech and text, the accuracy of road rage state recognition by different models was tested, as shown in Fig. 8.

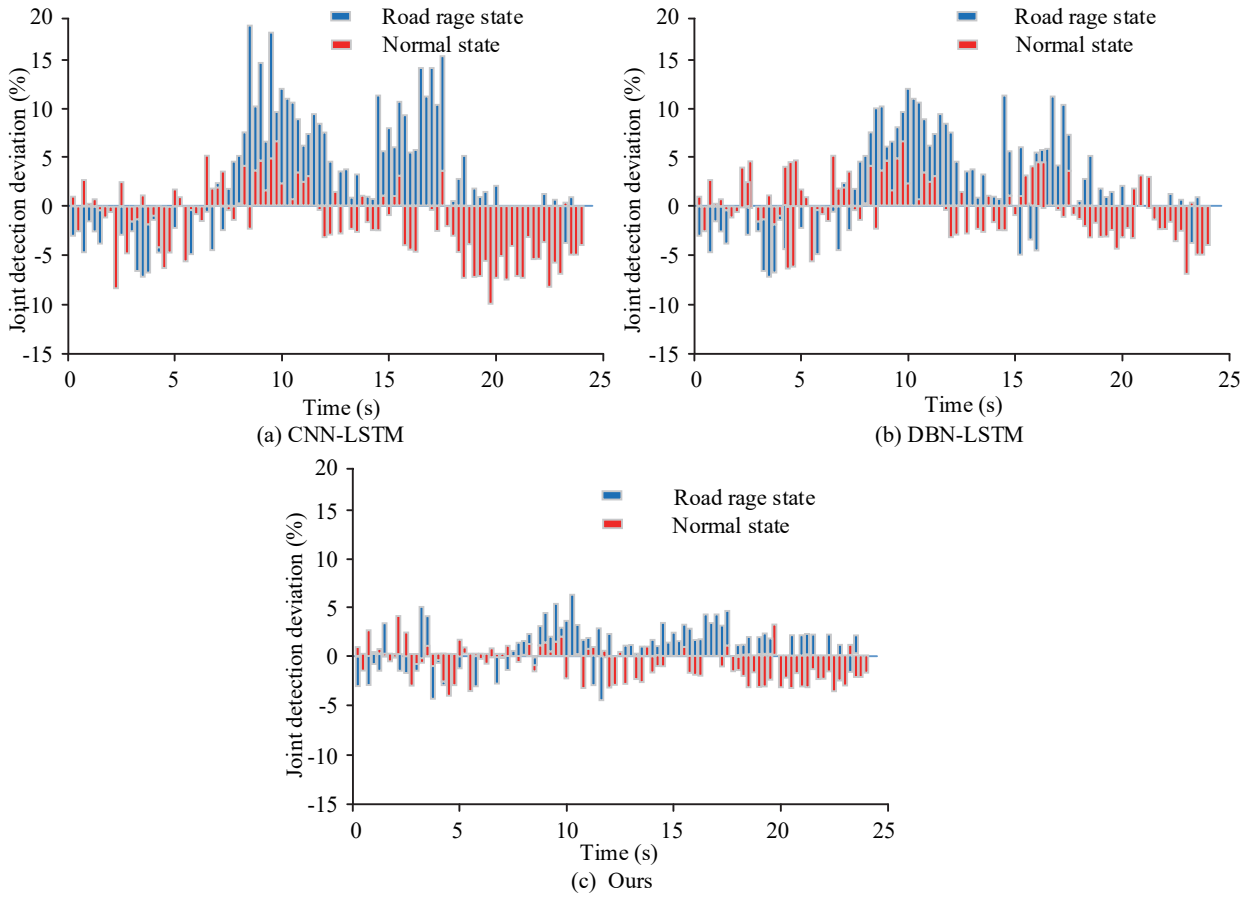


Figure 8 Deviation testing of road rage state detection using different models

Fig. 8a shows the road rage state detection results of the CNN-LSTM model. According to the test results, under multimodal data, the maximum deviation value of CNN-LSTM for road rage state detection was 19.2%, while the maximum deviation value for normal state detection was -9.8, indicating average overall performance. In the testing of the DBN-LSTM model in Fig. 8b, the deviation of the model for road rage state and normal state was controlled within the range of 11.2% and -5.2%, respectively, which was significantly better than CNN-LSTM. In the testing of the research model in Fig. 8c, the research model was more accurate in detecting two states, with the road rage state deviation controlled within the range of 5.3% and the normal state deviation controlled within the range of 4.7%, showing the best overall performance. It can be observed that the research model has the lowest deviation in road rage monitoring at 25 seconds, especially in the abnormal interval between 10 seconds and 20 seconds. This interval records the occurrence of dangerous behaviors such as verbal abuse and loud speech by drivers, and the research model accurately makes judgments, which can provide timely feedback to drivers to avoid increased road risks. Next, the precision, recall, and loss of different models were tested under multimodal data, as shown in Fig. 9.

Fig. 9a shows the precision test results. In 100 iterations, the maximum accuracy values of the research model, DBN-LSTM, and CNN-LSTM were 95.61%, 93.1%, and 92.3%, respectively. In the recall test of Figure 9 (b), the research model still performed the best, with a maximum recall value of 99.31%, while the maximum values of DBN-LSTM and CNN-LSTM were 98.1% and 93.7%, respectively. In the training loss test of Fig. 9c, the loss values of CNN-LSTM, DBN-LSTM, and the research model during iterative convergence were 1.231, 1.023, and 0.231, respectively. The model proposed by the research performs the best overall. Overall, the research model performs better due to the use of ISCSO to optimize PNN parameters, accurately capturing high-frequency speech

features, combined with LSTM to mine text temporal correlations, and enhancing feature complementarity through decision layer fusion. However, DBN-LSTM and CNN-LSTM have a single feature extraction and low fusion degree, resulting in lower accuracy, recall rate, and higher loss. Finally, the study selected a multimodal test set of 600 fused speech and text data, with road rage data accounting for 60%. The performance of the system was tested and verified during the testing process, as shown in Fig. 10.

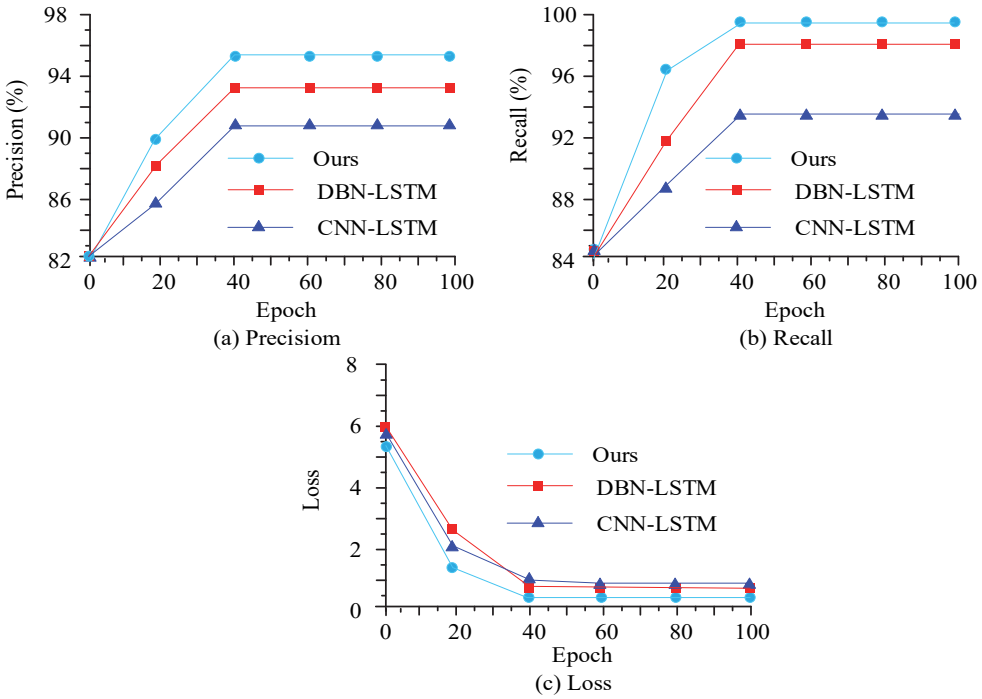


Figure 9 Performance test of multimodal data scene model detection

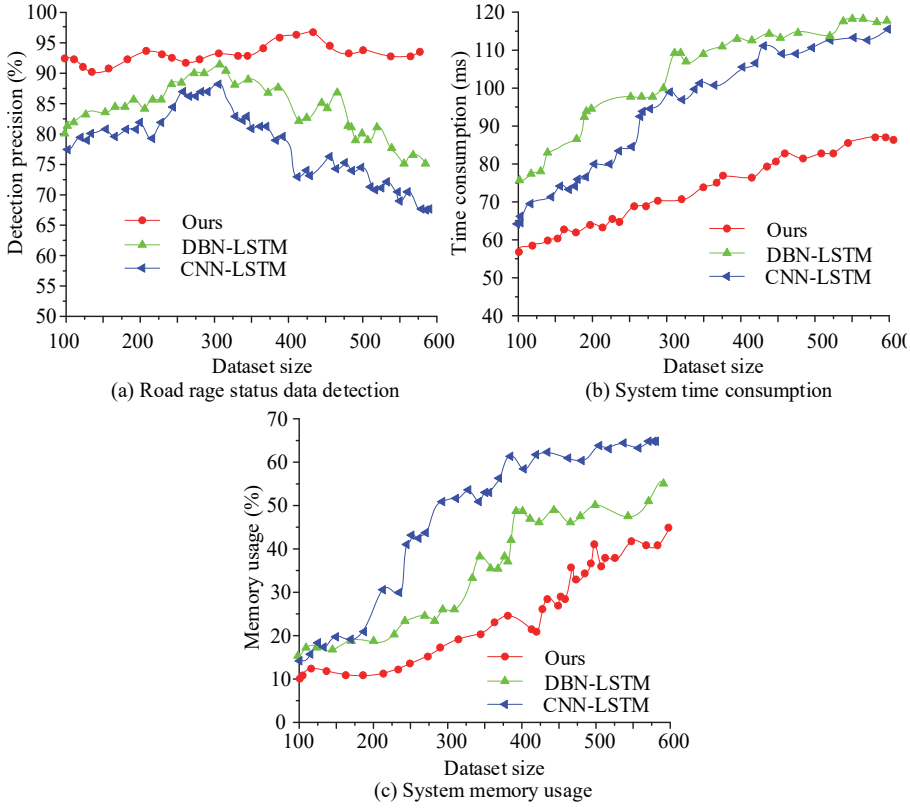


Figure 10 Comprehensive testing of multimodal scene system

Fig. 10a shows the precision results of data detection under different data scales, where the maximum precision of the research model was 96.82% and the minimum precision was 92.96. The minimum detection precision of DBN-LSTM and CNN-LSTM was 75.8% and 66.7%, respectively. In the system time test of Fig. 10b, as the data size increased, the time consumption of all three models gradually increased. The research model had a minimum

time of 58 ms and a maximum time of 89 ms, which was better than the other two models. Finally, in the system occupancy test shown in Fig. 10c, the minimum memory occupancy of the research model was only 10.06%, lower than the 16.87% and 15.47% of DBN-LSTM and CNN-LSTM, respectively. The method proposed by the research has excellent performance in traffic road rage detection, and overall performs better compared to similar technologies. According to the test results, an increase in driving scene detection data poses challenges to the system's data processing efficiency, recognition accuracy, and resource utilization. Especially in data containing provocation, insults, and abnormal noise, only the research model maintains high efficiency and recognition accuracy. This indicates that in some extreme scenarios, the research model can accurately make judgments on road rage, such as the appearance of slapping, insults, and multiple honking of the whistle. The research model can provide timely warnings and reminders to drivers to avoid dangerous driving based on recognition results, thereby reducing road driving risks. In addition, the study also introduced advanced models such as Transformer, BERT, and LSTM cross attention models (LSTM-CA) for robustness testing. The test is based on 30 independent experiments, using 5-fold cross validation to evaluate the model's generalization ability. The cross validation process is implemented using Python's Scikit learn library, and the Scipy library is used to calculate the mean, variance, and 95% confidence interval to ensure the reliability of the results. In terms of technical background, road rage detection needs to balance real-time performance and anti-interference, so we chose to compare the performance of different models in multiple scenarios within this framework.

Table 3 Robustness test results ($\bar{x} \pm s$, %)

Scene Type	DBN-LSTM	Transformer	BERT	CNN-LSTM	LSTM-CA	Ours
Congestion scenario	82.2 ± 1.8	85.5 ± 1.5	86.6 ± 1.7	87.7 ± 1.2	90.3 ± 0.9	94.1 ± 0.7
Expressway	84.5 ± 1.6	87.2 ± 1.3	86.7 ± 1.5	89.9 ± 1.1	92.6 ± 0.8	96.3 ± 0.6
Suburb	80.9 ± 2.0	84.1 ± 1.7	85.3 ± 1.7	86.6 ± 1.4	89.7 ± 1.0	93.5 ± 0.8
Rain and snow weather	78.3 ± 2.2	81.6 ± 1.9	82.3 ± 1.2	84.2 ± 1.6	87.5 ± 1.2	91.8 ± 0.9
Night lighting	83.4 ± 1.7	86.7 ± 1.4	87.4 ± 1.3	89.1 ± 1.3	91.9 ± 0.9	95.4 ± 0.7
Tunnel	79.6 ± 2.1	83.2 ± 1.8	84.7 ± 1.5	85.8 ± 1.5	89.0 ± 1.1	92.6 ± 0.8

According to the test in Tab. 3, after 5-fold cross validation, the performance ranking of each model remains stable, and the accuracy difference between adjacent models is 1.3%-3.8%, which meets the randomness requirements. And the research model has the best testing effect. For example, in congested scenarios, the highest accuracy of the research model is 94.1 ± 0.7, followed by LSTM-CA% and CNN-LSTM%, which are 90.3 ± 0.9% and 87.7 ± 1.2%, respectively. However, the overall performance of Transformer, BERT, and DBN-LSTM is average. The research model fully integrates speech and text data, while enhancing the extraction of noisy data. The test results are significantly better than other models. Although LSTM-CA enhances the attention to anomalous features, it has poor adaptability to noise. In addition, although Transformer and LSTM-CA have added attention mechanisms, their performance stability is inferior to the research model in high noise scenarios such as rainy and snowy days, with variances of 1.9 and 1.2, respectively. It can be seen that research techniques have good robustness in multiple scenarios.

5 DISCUSSION

Traffic accidents caused by road rage occur frequently and seriously threaten road safety. At present, the traditional single mode detection technology has problems such as single information source and insufficient recognition rate. A multimodal detection technique combining speech and text is proposed to improve the detection of road rage in road traffic scenes and reduce traffic risks.

In the single modal performance verification, the ISCSO-PNN model performed excellently. In simple scenarios, CNN has 1 unrecognized and 4 misjudgments, DBN has 3 misjudgments, and ISCSO-PNN has no unrecognized or misjudgments. This is due to ISCSO optimizing PNN parameters to avoid overfitting and inter class boundary blurring, enhancing the capture of high-frequency features of road rage speech, while CNN and DBN are not optimized for speech noise, which interferes with feature extraction. In text scenarios, the improved LSTM achieved average recognition accuracies of 88.6%, 81.0%, 87.6%, and 84.2% in noun, verb, adjective and special vocabulary data tests, respectively, all higher than LSTM and TextCNN. This is because it effectively captures long-term dependencies of text sequences through gating mechanisms and enhances the recognition of emotional vocabulary by combining it with an emotional dictionary. However, similar technologies lack emotional level feedback on emotional vocabulary, making it difficult to accurately determine whether drivers are at risk of road rage. In addition, the advantages of studying models are more significant in multimodal detection. Compared with DBN-LSTM and CNN-LSTM, its detection deviation for road rage and normal states is smaller, controlled within 5.3% and 4.7% respectively; The maximum accuracy is 95.61% and the recall is 99.31%, both higher than the comparison model; The training loss value is 0.231, much lower than the other two values of 1.023 and 1.231. This is because the model integrates speech and text data, with ISCSO-PNN accurately capturing high-frequency features of speech, LSTM mining temporal correlations of text, and decision layer fusion enhancing feature complementarity. In contrast, the feature extraction of the comparative model is single and the fusion degree is low. Compared with the aggressive driving detection scheme studied by Aljagoub et al. [3], the multimodal fusion technology has stronger adaptability in complex scenarios, obvious advantages in accuracy and recall, and can more accurately identify road rage behavior.

In terms of actual vehicle deployment, this technology collects voice through a microphone and transcribes text through an ASR engine, with low hardware requirements for the car. Compared with visual monitoring and heart rate

monitoring, the research technology is suitable for most in car systems, and its data timeliness processing is excellent. Compared with visual and heart rate monitoring technologies, the research technology has a low memory usage of only 10%, while video and heart rate detection technologies both have a usage rate of over 40%, with the highest computational efficiency among the three and the lowest deployment cost. In addition, in terms of driver privacy security, the research technology processes data through local edge computing to avoid uploading original data. If the original voice and text data are not uploaded to the cloud, only the desensitized feature parameters are transmitted to reduce the risk of leakage. The research technology also considers the issue of false alarms. The research technology adopts an adaptive weighting method based on decision layer fusion, which provides effective speech modality confidence based on the frequency of driver's voice, text, and abnormal data. Only when it reaches 0.9 will a warning occur, avoiding the problem of false alarms in traditional technology. In addition, a survey of 1200 drivers showed that 92.5% recognized the value of the model for safety and 90.3% were willing to deploy it due to its low false alarm rate. More drivers believe that technology does not provide video monitoring of driver privacy information, and that technology deployment requires low requirements, low costs, and weak risk interference to drivers, making it more suitable for the needs of most drivers.

It can be seen that the technology proposed by the research institute has excellent performance effects in practical scenarios. And it is superior to similar technologies in deployment, security, privacy, and feasibility, meeting the usage requirements of most drivers and providing technical support for road traffic safety.

6 CONCLUSION

Road rage driving is an important issue in the field of traffic safety. Drivers in a state of road rage are prone to losing their rationality, which has a serious impact on traffic safety. Therefore, this paper presented a multimodal approach for detecting road rage, combining ISCSO-optimized PNNs for speech and LSTMs for text, integrated through decision-level fusion. The system demonstrated significant improvements in precision, recall, and efficiency compared with CNN-LSTM and DBN-LSTM baselines, achieving accuracy up to 95.6% and recall up to 99.3%. Experiments confirmed robustness under noisy conditions and efficiency suitable for real-time use. The multimodal road rage detection model studied performs well, but has limitations as it only uses speech and text data. Future work needs to integrate multi-source data such as facial expressions, body posture, and vehicle telemetry, conduct transformer based multimodal architecture benchmark testing, and conduct on-site testing in real driving environments to improve the framework and make it a practical tool for improving road safety through real-time road rage detection.

7 REFERENCES

[1] Wang, D. L., Ding, A., Chen, G. L., & Zhang, L. (2023). A combined genetic algorithm and A* search algorithm for the electric vehicle routing problem with time windows. *Advances in Production Engineering & Management*, *18*(4), 403-416. https://doi.org/10.14743/apem2023.4.481

[2] Wan, P., Jing, X., Lu, S., & Yan, L. (2023). Impact of temperament types and anger intensity on drivers' EEG power spectrum and sample entropy: an on-road evaluation toward road rage warning. *Tehnički vjesnik*, *30*(4), 1055-1067. https://doi.org/10.17559/TV-20221021054632

[3] Aljagoub, D., Ardeshir, F., & Karakurt, A. (2023). The Plague of Aggressive Driving: Definitions, Causes, Severity, Consequences, and Solutions. *Open Journal of Safety Science and Technology*, *13*(3), 132-151. https://doi.org/10.4236/ojsst.2023.133007

[4] Sar, I., Routray, A., & Mahanty, B. (2023). A review on existing technologies for the identification and measurement of abnormal driving. *International Journal of Intelligent Transportation Systems Research*, *21*(1), 159-177. https://doi.org/10.1007/s13177-023-00343-7

[5] Arumugam, S. & Bhargavi, R. (2023). Road rage and aggressive driving behaviour detection in usage-based insurance using machine learning. *International Journal of Software Innovation (IJSI)*, *11*(1), 2-29. https://doi.org/10.4018/IJSI.319314

[6] Hmidi, N., Afdhal, R., & Hamdi, M. (2024). Emotion estimation of people wearing masks using machine learning. *International Journal of Computers Communications & Control*, *19*(1), 5363-5365. https://doi.org/10.15837/ijccc.2024.1.5363

[7] Liu, P. (2024). Machines meet humans on the social road: Risk implications. *Risk analysis*, *44*(7), 1539-1548. https://doi.org/10.1111/risa.14255

[8] Yan, L. X., Jia, L., Guo, J. H., & Lu, S. (2022). A simulation study on the identification of eco-driving behaviour. *International Journal of Simulation Modelling*, *21*(3), 489-500. https://doi.org/10.2507/IJSIMM21-3-CO11

[9] Zhang, Y., Song, P., Song, Q., & Li, Q. (2023). A Novel Point Cloud Compression Algorithm for Vehicle Recognition Using Boundary Extraction. *Tehnički vjesnik*, *30*(6), 1899-1910. https://doi.org/10.17559/TV-20230507000612

[10] Hasan, M. A. (2023). Facial human emotion recognition by using YOLO faces detection algorithm. *JOINCS (Journal of Informatics, Network, and Computer Science)*, *6*(2), 32-38. https://doi.org/10.21070/joincs.v6i2.1629

[11] Surana, A., Rathod, M., Gite, S., Patil, S., & Kotecha, K. (2024). An audio-based anger detection algorithm using a hybrid artificial neural network and fuzzy logic model. *Multimedia Tools and Applications*, *83*(13), 38909-38929. https://doi.org/10.1007/s11042-023-16815-7

[12] Halim, Z., Sulaiman, M., Waqas, M., & Aydın, D. (2023). Deep neural network-based identification of driving risk utilizing driver dependent vehicle driving features: A scheme for critical infrastructure protection. *Journal of Ambient Intelligence and Humanized Computing*, *14*(9), 11747-11765. https://doi.org/10.1007/s12652-022-03734-y

[13] Yang, L., Yang, H., Hu, B. B., & Wang, Y. (2023). A robust driver emotion recognition method based on high-purity feature separation. *IEEE Transactions on Intelligent Transportation Systems*, *24*(12), 15092-15104. https://doi.org/10.1109/TITS.2023.3304128

[14] Gong, P., Wang, P., Zhou, Y., & Wen, X. T. (2024). fac-net: A temporal-frequential attentional convolutional network for driver drowsiness recognition with single-channel eeg. *IEEE Transactions on Intelligent Transportation Systems*, *25*(7), 7004-7016. https://doi.org/10.1109/TITS.2023.3347075

[15] Mou, L., Zhao, Y., Zhou, C., & Nakisa, B. (2023). Driver emotion recognition with a hybrid attentional multimodal fusion framework. *IEEE Transactions on Affective Computing*, *14*(4), 2970-2981. https://doi.org/10.1109/TAFFC.2023.3250460

[16] Katual, J. & Kaul, A. (2024). Optimized Ensemble Machine Learning Approach for Emotion Detection from Thermal Images. *International Journal of Pattern Recognition and Artificial Intelligence*, *38*(02), 2451002. https://doi.org/10.1142/S0218001424510029

[17] Deshmukh, S. & Gupta, P. (2024). Application of probabilistic neural network for speech emotion recognition. *International Journal of Speech Technology*, *27*(1), 19-28. https://doi.org/10.1007/s10772-023-10037-w

[18] Banzon, A. M., Beever, J., & Taub, M. (2023). Facial expression recognition in classrooms: Ethical considerations and proposed guidelines for affect detection in educational settings. *IEEE Transactions on Affective Computing*, *15*(1), 93-104. https://doi.org/10.1109/TAFFC.2023.3275624

[19] Bethge, D., Coelho, L. F., Kosch, T., Murugaboopathy, S., Zadow, U., Schmidt, A., & Grosse-Puppendahl, T. (2023). Technical design space analysis for unobtrusive driver emotion assessment using multi-domain context. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, *6*(4), 1-30. https://doi.org/10.1145/3569466

[20] Li, J., Wang, X., Lv, G., & Zeng, Z. (2023). GA2MIF: Graph and attention based two-stage multi-source information fusion for conversational emotion detection. *IEEE Transactions on affective computing*, *15*(1), 130-143. https://doi.org/10.1109/TAFFC.2023.3261279

[21] Hu, R. & Huang, P. (2024). Autonomous Driving Decision-Making Based on an Improved Actor-Critic Algorithm. *Studies in Informatics and Control*, *33*(4), 37-50. https://doi.org/10.24846/v33i4y202404

[22] Hourri, S. (2024). Empowering Speaker Verification with Deep Convolutional Neural Network Vectors. *Studies in Informatics and Control*, *33*(2), 97-107. https://doi.org/10.24846/v33i2y202409

[23] Wu, Y., Daoudi, M., & Amad, A. (2023). Transformer-based self-supervised multimodal representation learning for wearable emotion recognition. *IEEE Transactions on Affective Computing*, *15*(1), 157-172. https://doi.org/10.1109/TAFFC.2023.3263907

[24] Xuan, W. & Deng, M. (2023). Logistics service quality sentiment analysis with deeper attention LSTM model with aspect embedding. *Tehnički vjesnik*, *30*(2), 634-641. https://doi.org/10.17559/TV-20221018031450

[25] Renjith, P. N., Balasubramani, S., & Ramesh, K. (2025). An Initial Risk Assessment for Multimodal with LSTM-Based Trust Evaluation Framework for Autonomous Vehicle Security. *SN Computer Science*, *6*(2), 1-15. https://doi.org/10.1007/s42979-025-03703-0

[26] Karas, V., Schuller, D. M., & Schuller, B. W. (2023). Audiovisual affect recognition for autonomous vehicles: Applications and future agendas. *IEEE Transactions on Intelligent Transportation Systems*, *25*(6), 4918-4932. https://doi.org/10.1109/TITS.2023.3333749

Contact information:

Enlin XIE
(Corresponding author)
Xiangsihu College of Guangxi Minzu University,
Nanning 530000, P. R. China
E-mail: foolishxel@163.com

Yiliu HUANG
Teachers Training Center of Guangxi Zhuang Autonomous Region Nanning,
530018, China
E-mail: Hyiliu20092005@163.com