RESEARCH ARTICLE

Deep learning-based bimodal speech and facial expression recognition of miners' unsafe emotions

Ying Lu[1,2,3]*, **Zihao Zhao**[1], **Liang Yan**[4]*, **Xinyv Shi**[1]

1 School of Resource and Environmental Engineering, Wuhan University of Science and Technology, Wuhan, Hubei, China, **2** Hubei Industrial Safety Engineering Technology Research Center, Wuhan, Hubei, China, **3** School of Management, Fudan University, Shanghai, China, **4** Hubei Chemical Safety Association, Wuhan, Hubei, China

* luying@wust.edu.cn (YL); 2516542858@qq.com (LY)

OPEN ACCESS

Citation: Lu Y, Zhao Z, Yan L, Shi X (2026) Deep learning-based bimodal speech and facial expression recognition of miners' unsafe emotions. PLoS One 21(5): e0348906. https://doi.org/10.1371/journal.pone.0348906

Editor: Bilal Alatas, Firat Universitesi, TÜRKIYE

Received: November 7, 2025

Accepted: April 21, 2026

Published: May 15, 2026

Data availability statement: The dataset used in this study includes both publicly available data and a self-constructed dataset. Anger and fear emotion data were obtained from the publicly available CASIA database, which can be accessed directly from the original source. Anxiety, irritability, and ennui data were collected as part of a self-constructed dataset involving human participants. Due to informed

Abstract

Under the influence of unsafe emotions, miners' ability to perceive risks is hindered, which can easily lead to decision-making errors and safety accidents. To recognize unsafe emotions exhibited by miners during operations, this study proposes a deep learning-based bimodal framework that integrates speech and facial expression features. A convolutional neural network (CNN) combined with a bidirectional long short-term memory (Bi-LSTM) network is employed to model local spectral patterns and temporal dependencies in speech signals, and ShuffleNet-V2 is used to capture deep facial features. In addition, three feature enhancement strategies are proposed to improve the generalization ability of the model. By constructing a dataset containing five categories of miners' unsafe emotions for network training, the model achieves a mean recognition accuracy of 85.56%. Furthermore, we conducted a preliminary field test of the bimodal model in a real mining environment. The results provide preliminary evidence of its potential applicability in real-world mining conditions.

1 Introduction

Mineral resources are the cornerstone of stable energy supply and sustainable economic development [1–3]. Nevertheless, the persistent occurrence of safety accidents in the mining sector constitutes a critical global challenge. In China, statistical data has indicated a 23.7% increase in the mortality rate per million tons in coal mines during 2023, thereby emphasising the necessity for more proactive safety interventions [4]. Recent interdisciplinary studies have demonstrated that miners' unsafe behaviour is significantly driven by their emotional states [5–6]. Specifically, negative emotions act as a root cause for psychological instability, leading to impaired risk perception and decision-making errors [7–8]. In the high-pressure, harsh environments of underground mines, characterised by noise, dust, and

consent agreements and ethical requirements, these data cannot be shared publicly. Data are available from the Academic Committee of the School of Resource and Environmental Engineering, Wuhan University of Science and Technology (contact: Dr. Jun Han, Chair; Email: hanjun@wust.edu.cn) for researchers who meet the criteria for access to confidential data.

Funding: This work was supported by the Special Project of Safety Production of the Hubei Emergency Management Department (No. SJZX20230907). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

isolation, miners are highly susceptible to anxiety, irritability, and burnout [9–12]. As indicated by previous research, miners experiencing negative emotional states have been shown to be 17 times more likely to incur injuries than those in a positive emotional state [6]. Consequently, the real-time recognition of miners' unsafe emotions is imperative for the prediction of potential safety breaches and the prevention of accidents.

Against this background, the need for effective, objective, and scalable perception technologies to monitor miners' emotional states has become increasingly evident. Recent studies have demonstrated the strong representational capacity of deep learning models across diverse industrial perception tasks, including medical image segmentation [13], vision transformer-based visual understanding [14], and CNN-driven speech recognition under low-resource conditions [15]. By contrast, the traditional safety inspection method for miners' emotion still relies heavily on manual observation, which is inherently inefficient and subjective [16].

Therefore, an increasing number of deep learning-based emotion recognition models have been applied in safety-critical fields and industrial settings, including mining operations. Early studies predominantly focused on facial expression recognition, leading to the development of miner-specific models such as miniXception [17]. However, these unimodal approaches are highly sensitive to common disturbances in underground tunnels, including dust interference, insufficient lighting, and physical occlusions [18–19]. Alternative methods employing physiological sensors, such as electroencephalogram (EEG) [20–22] or electrodermal activity (EDA) [23–24], have the capacity to provide objective data. However, these methods are costly, uncomfortable for prolonged use, and difficult to deploy at scale in harsh mining operations [25].

To overcome the limitations of unimodal emotion recognition, recent studies have increasingly turned to multimodal frameworks that integrate complementary information sources [26–27]. However, extant multimodal frameworks generally employ power-intensive Transformer architectures [28] or rely on increasingly complex model designs to enhance prediction performance [29]. While such architectures demonstrate superior representation capability, they typically incur substantial computational and memory overhead, leading to high latency and energy consumption. Several recent studies have reported that Transformer-based multimodal emotion recognition models require significantly larger parameter sizes and longer inference times compared with lightweight CNN-based alternatives, which limits their deployment on edge devices and real-time monitoring systems [30–31]. In the mining environment, such computational demands pose a significant practical obstacle due to the severe constraints on computing resources [25]. Consequently, there is a pressing need for lightweight multimodal architectures that can achieve a favorable balance between recognition accuracy and computational efficiency, ensuring stable operation under harsh mining conditions.

Furthermore, while existing studies provide valuable insights into multimodal emotion recognition performance under controlled conditions, most publicly available

datasets are collected in laboratory or semi-controlled environments, which fail to reflect the complex and dynamic characteristics of real-world mining scenarios [25,27]. Variations in background noise, illumination changes, visual occlusions, and individual behavioral differences can substantially degrade model performance when transferred to real-world settings [19,32–34]. Therefore, field experiments conducted in actual mining environments are essential for objectively assessing model robustness and practical applicability. Moreover, systematically analyzing the influence of environmental factors, such as sound intensity, distance, and visual occlusion, on recognition accuracy can provide critical evidence for targeted model optimization and deployment strategies.

A review of the current literature reveals three main research gaps. First, existing miner emotion recognition models primarily rely on facial expressions, making them vulnerable to environmental disturbances. Second, many multimodal frameworks suffer from large parameter sizes and high computational costs, limiting their portability. Third, current approaches lack the application exploration under real-world mining conditions, and existing datasets fail to reflect emotional variability.

In order to explicitly address the above research gaps, this paper proposes a bimodal deep learning model that has been specifically optimised for mine safety. A hybrid Convolutional Neural Network (CNN) and Bidirectional Long Short-Term Memory (Bi-LSTM) is employed to extract time-domain and frequency-domain features from speech, while an enhanced ShuffleNet-V2 is utilised for facial expression analysis. In contrast to conventional large-scale networks, ShuffleNet-V2 was selected for its well-structured architecture, which was developed specifically for mobile and edge devices, thereby achieving a superior balance between accuracy and scalability.

The main contributions of this study are summarized as follows:

(1) A deep learning-based bimodal framework is proposed to recognize miners' unsafe emotions by jointly modeling speech and facial expression information.

(2) Three feature enhancement strategies, including frequency masking, coordinate attention, and squeeze-excitation mechanisms, are introduced to improve model generalization under harsh mining conditions.

(3) A dedicated dataset of miners' unsafe emotions is constructed, covering five emotion categories relevant to mining safety. The proposed model is validated through numerical experiments, indoor simulations, and real mining field tests, which provides preliminary evidence of its potential applicability in mine operation scenarios.

The remainder of this study is organized as follows: Sect 2 describes the methodology, including dataset construction, model architecture, and experimental procedure. Sect 3 presents the results from numerical, indoor, and field experiments. Sect 4 discusses the model's performance and environmental influences. Finally, Sect 5 concludes the paper and outlines future research directions.

2 Methodology

The framework of the bimodal unsafe emotion recognition model proposed in this paper is shown in Fig 1, which firstly constructs speech and facial expression datasets separately, and then performs feature extraction on the speech and facial expression data respectively under the premise of guaranteeing the feature integrity, and finally outputs the emotion classification after feature fusion. We propose three types of feature enhancement methods to optimize the model. Specifically, for speech modality, MFCC features augmented with frequency masking are extracted and processed using a CNN–BiLSTM network with attention. For facial modality, LBP and WLD features are fused and analyzed using an improved ShuffleNet-V2 network with a squeeze-excitation mechanism. Through numerical experiments, indoor experiments and field experiments, the effectiveness of the model in different scenarios was preliminarily explored, and finally a bimodal miners' unsafe emotion recognition system was constructed.

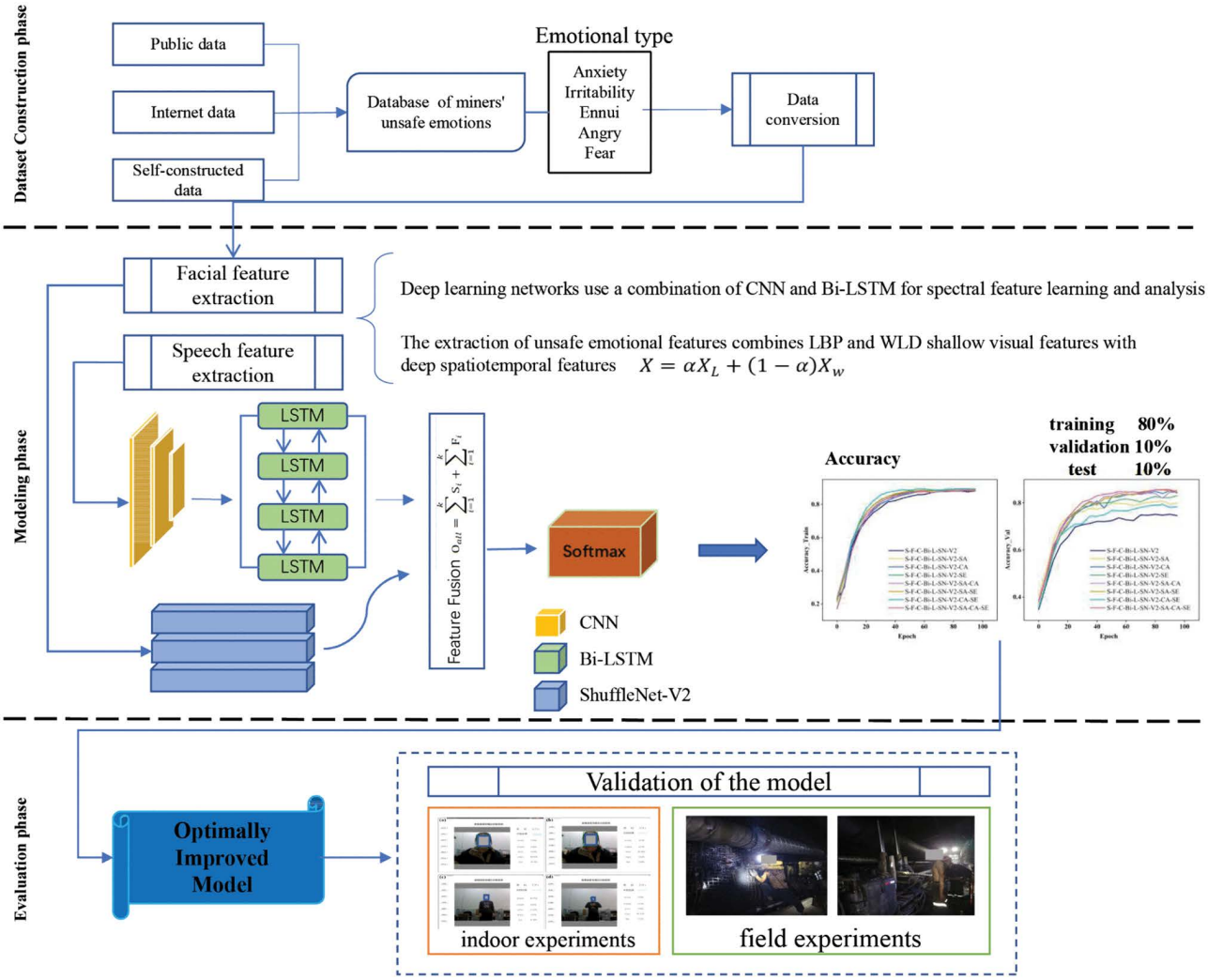

Fig 1. The main flow of the unsafe emotion recognition model. The individuals shown in this figure have given their written permission to publish their photographs under the CC BY 4.0 license.

https://doi.org/10.1371/journal.pone.0348906.g001

2.1 Ethics statement

This study was conducted in accordance with Item 32 of the Ethical Review Measures for Human Life Science and Medical Research issued by the National Health Commission of the People's Republic of China (effective 27 February 2023) and qualifies for exemption from formal ethical review. The research involves only the collection of voice and facial expression data, poses no physical or psychological risk, and does not interfere with normal work activities.

All participants were fully informed of the study objectives, procedures, and data usage and provided written informed consent. Participation was voluntary, and participants could withdraw at any stage. The design and conduct of this study strictly adhere to the requirements stipulated for exempt research within the aforementioned regulation, particularly concerning respect for participant rights (as applicable to the data source), privacy protection, and data security.

2.2 Dataset

Most publicly available datasets are built on base emotions such as anger, surprise, frustration, happiness, fear and sadness, while it is difficult to accurately represent the categories of unsafe emotions such as anxiety and irritability, which have been mentioned in the study of miners' emotions. Therefore, the dataset in this paper was collected in a variety of ways. Among them, the two categories of miners' emotions, anger and fear, were collected from the CASIA database of the Chinese Academy of Sciences with 703 speech signals and 824 expression images. For the three emotion categories of anxiety, irritability and ennui, due to the difficulty of obtaining them from public databases, this study also developed an emotion dataset. The collected data were independently annotated by two co-authors following a predefined annotation protocol, without access to each other's labeling results. To evaluate the consistency of manual annotations, inter-rater reliability was measured using Cohen's Kappa coefficient for both voice and image modalities. The Cohen's Kappa coefficient for the voice modality was 0.712, and for the image modality was 0.719, indicating substantial agreement between the two annotators. The confusion matrices of the annotations for the voice and image modalities are presented in Tables 1 and 2, respectively.

For the annotations with differences, a consensus was reached through group discussion and the issues were resolved. A database of miners' unsafe emotions containing 1,736 voice signals and 1,981 facial expression images was finally obtained after the labeling of emotion categories, and the detailed information of unsafe emotions in each category is shown in Table 3. The dataset was randomly stratified and divided into training, validation and test sets with a ratio of 80%, 10% and 10%, respectively.

Table 1. Confusion matrix of speech emotion categories based on manual annotation.

Voice	Anxiety	Irritability	Ennui	Total (Annotator 1)
Anxiety	312	48	42	402
Irritability	45	318	39	402
Ennui	36	33	397	466
Total (Annotator 2)	393	399	478	1270

https://doi.org/10.1371/journal.pone.0348906.t001

Table 2. Confusion matrix of facial emotion categories based on manual annotation.

Image	Anxiety	Irritability	Ennui	Total (Annotator 1)
Anxiety	358	52	46	456
Irritability	48	331	42	421
Ennui	39	35	460	534
Total (Annotator 2)	445	418	548	1411

https://doi.org/10.1371/journal.pone.0348906.t002

Table 3. Database details of miners' unsafe emotions.

	Category	Number of voices	Number of images	Total
Emotional type	Anxiety	315	362	Voice signal: 1736 Expression image: 1981
	Irritability	318	334	
	Ennui	400	461	
	Anger	390	497	
	Fear	313	327	

https://doi.org/10.1371/journal.pone.0348906.t003

All data were obtained through legal channels or generated through observation without interfering with normal work activities. Data collection and analysis strictly complied with the terms and conditions of the respective data sources and adhered to Item 32 of the Ethical Review Measures for Human Life Science and Medical Research issued by the National Health Commission of the People's Republic of China (effective 27 February 2023). Personal privacy was fully protected throughout the study.

2.3 Modeling of bimodal recognition of miners' unsafe emotions

2.3.1 Speech emotion recognition. Speech emotion recognition consists of two parts, first extracting audio features from audio data and then analyzing it using deep learning networks.

Audio feature extraction is the key step in speech emotion recognition. Since emotions can be expressed through the speaker's pronunciation, pitch, intensity, and many other acoustic features, proper selection of audio features has a significant impact on the performance of speech emotion recognition. Mel Frequency Cepstrum Coefficient (MFCC) was proposed by Davis [35], and is a widely used feature in the field of speech recognition.The MFCC is a powerful audio feature that captures the speech signal's unique features for more accurate emotion classification and is the core of various speech emotion recognition studies. The audio data is first pre-emphasized, then the signal is processed in frames, multiplying each frame by a Hamming window to increase the continuity of the left and right ends, the frequency domain signal of each frame is obtained after FFT, the logarithmic spectrum is computed, and a bank of 64 Mel filters ranging from 20 to 8,000 Hz is applied to the frequency domain signal, the logarithmic operation is performed on the outputs of the filters, the dynamic range is suppressed, and finally the output of the filter is Arrange the output of the previous step to form the entire Mel spectrogram of the speech signal.

Before inputting the network, this paper employs the frequency masking method from SpecAugment to augment the spectrogram. Since it operates directly on the spectrogram without requiring additional audio data, this method is highly straightforward and computationally inexpensive. Furthermore, during augmentation, it can randomly remove certain frequency channels from the spectrogram of speech signals to simulate noise or damage. This enables the miner speech recognition module to better learn speech signal features, thereby enhancing model performance.

CNN + Bi-LSTM model is a deep learning model commonly used for emotion recognition tasks. It combines CNN and Bi-LSTM to fully capture contextual information in the text and model local features. Bi-LSTM is able to handle longer text sequences and maintains the ability to model long-distance dependencies. Combining these two models can simultaneously consider the spatio-temporal features of the miner's emotional state and improve the model's ability to understand the miner's emotional state.The CNN structure used in this paper is shown in Table 4. Although CNN and Bi-LSTM perform well in image and sequence data processing, it may not be possible to fully capture the features of all emotional states using only CNN + Bi-LSTM. Different emotional states may exhibit different patterns and differences. Therefore, the introduction of Attention module in the miner's emotion classification model can be used to assign weights according to the different impacts of input features on the output results, and improve the feature extraction ability of the network model by differentiating the weights.

Table 4. Parameter settings of CNN.

	Layer	KSize	Stride	Repeat
Conv	Convolution	5 × 5	2	3
	MaxPool	2 × 2	2	
	BN			
	ReLU			
	Dropout			

https://doi.org/10.1371/journal.pone.0348906.t004

The Attention mechanism is modeled after the efficient information processing capability of the human visual system, which can quickly locate and focus on the key areas of the observed objects, and assign higher weights to the important targets, so as to achieve efficient information screening and recognition in complex scenes. Currently, this mechanism has been widely used in machine translation, text summarization, speech recognition and other tasks, and has become an indispensable and efficient module in deep learning models, especially in dealing with long sequence dependencies, cross-modal alignment and complex scene understanding and other scenarios, showing significant advantages [36]. Hou et al. [37] proposed a novel efficient coordinate attention mechanism, which integrates both horizontal and vertical positional information from feature maps into channel attention. This innovation enables mobile networks to capture extensive spatial information without excessive computational overhead. The proposed attention mechanism effectively captures long-range dependencies in speech features, dynamically adjusting the importance of different positions within feature maps. By assigning higher weights to emotion-related information, it achieves more precise target localization and enhances the contribution of emotional components to final recognition outcomes.

2.3.2 Facial emotion recognition. Facial emotion recognition comprises two main components: feature extraction and network analysis. For feature extraction, we employ a fusion of Local Binary Patterns (LBP) and Weber Local Descriptor (WLD) algorithms. LBP captures the edge direction information of images, while WLD focuses on the local intensity information. After extracting LBP and WLD features separately, we obtain the combined features by setting weighted fusion coefficients, as described in (1). These features are then input into the ShuffleNet-V2 network for analysis.

$$X = \alpha X_L + (1 - \alpha) X_w \tag{1}$$

where X_L is the LBP feature, X_w is the WLD feature, and X is the hybrid feature

ShuffleNet V2 is an efficient neural network designed for high computational performance, accuracy, and lightweight deployment. The ShuffleNet-V2 network reduces the computational complexity of the model by stacking the compact operator deep convolution, and the convolution blocks stage2, stage3, and stage4 inside the network are composed of two parts: the first part is the first DownBlock in each stage. Since this module needs to downsample and enhance the dimension, the input information is directly copied into two copies, and the feature extraction is performed through the two branches of Branch1 and Branch2, respectively, after which the extracted feature information is spliced along the channel dimension, and finally the channels of the stacked feature maps are reordered through the Channel Shuffle to realize the feature fusion between the two branches. of the feature fusion. The second part is a number of BasicBlock in each stage, firstly, the number of channels of the input feature map is equally divided into two branches through Channel Spilt, the first path carries out residual linking to retain the original state, and the other branch realizes the extraction of feature information. Finally, the two branches are subjected to Channel Dimension Split as well as Channel Shuffle operation, so that the shallow feature information is effectively transferred to the deep network to achieve the effect of feature reuse.

To optimize the ShuffleNet-V2 network, we insert a squeeze-excitation (SE) mechanism at the end of the ShuffleNet unit. This enhancement addresses the issue in the bimodal emotion recognition model, where the facial emotion extraction module may not fully utilize the local feature information of images. The SE mechanism strengthens the directional features of facial expressions across five categories of unsafe emotions, thereby mitigating the problem of model overfitting.The overall network structure of the final ShuffleNet-V2 is shown in Table 5.

2.3.3 Feature fusion. The speech and facial expression features of miners' unsafe emotions are inherently sequential and often exhibit correlations. Therefore, during model construction, feature-level fusion of speech and facial expression features is employed to achieve multimodal information complementarity, leading to more accurate emotion assessment. The concatenation operation effectively preserves the distinct unsafe emotion information across different modalities [38]. Assuming that the speech feature vector group $O_{speech} = [S_1, S_2, \ldots S_k]$ and the expression feature vector group $O_{face} = [F_1, F_2, \ldots F_k]$ are obtained after feature extraction, then the fused feature

Table 5. Network structure of ShuffleNet-V2.

Layer	KSize	Stride	Repeat
Conv1	3 × 3	2	1
MaxPool	3 × 3	2	1
Stage2		2	1
		1	3
Stage3		2	1
		1	7
Stage4		2	1
		1	3
SENet			1
Conv5	1 × 1	1	1
GlobalPool	7 × 7		

https://doi.org/10.1371/journal.pone.0348906.t005

vector group $O_{all} = [A_1, A_2, \ldots A_k]$ is obtained after concatenation, and finally the classification result is obtained by inputting into the softmax layer as shown in (2).

$$O_{all} = [O_{speech}; O_{face}] \tag{2}$$

where [;] denotes the concatenation operation.

2.4 Experimental program

We designed a series of numerical experiments to validate the performance of our proposed speech feature extraction method, facial expression feature extraction method, and the bimodal miner's unsafe emotion recognition model based on the feature layer fusion strategy, as well as to test the generalization ability of the model through two field experiments.

2.4.1 Numerical experiments. The relevant configuration of the experimental computer is Intel Core i5-9300H CPU, 16GB RAM, NVIDIA GeForce GTX 1660Ti, the experimental environment is Windows, the programming language is Python, and the deep learning framework is Pytorch. Our models were trained using a mini-batch size of 16 and categorical cross-entropy as a loss function. The Adam optimization algorithm for the initial learning rate of 0.001 and 100 epochs. To avoid the effect of overfitting, we employed an early stopping mechanism, which terminates training if the validation loss fails to improve for 10 consecutive epochs.

In order to verify the effectiveness of the enhancement strategies for speech and expression unsafe emotion features in the bimodal miner's unsafe emotion recognition model, seven different improvement strategies are formulated for numerical experiments by combining three different types of network modules using S-F-C-Bi-L-SN-V2 as the baseline model, as shown in Table 6. To improve the robustness and generalizability of the experimental results, repeated random subsampling validation was conducted. Specifically, the dataset was randomly stratified and divided into training, validation, and test sets with a ratio of 80%, 10%, and 10% for five independent runs. Each model (baseline model and seven improved models) was implemented and evaluated under the same conditions, including identical hardware, software, hyperparameters, and the same five sets of stratified data partitions, and the final results were reported as the mean±-standard deviation across all runs.

2.4.2 Indoor and field validation experiments. In order to further verify the effectiveness of the improved optimal model for miners' common unsafe emotion recognition, indoor experiments and field experiments are set up to analyze the performance of the model in different scenarios. Meanwhile, to facilitate the experiments, a prototype bimodal miner's unsafe emotion recognition system for preliminary practical tests was developed using

Table 6. Details of the numerical experimental program for different improvement strategies.

	Improvement program	Database	Evaluation index
Baseline model	S-F-C-Bi-L-SN-V2	Database of miners' unsafe emotions	Accuracy, F1-Score, Loss Value
Improved models	S-F-C-Bi-L-SN-V2 + CA		
	S-F-C-Bi-L-SN-V2 + SA		
	S-F-C-Bi-L-SN-V2 + SE		
	S-F-C-Bi-L-SN-V2 + SA + CA		
	S-F-C-Bi-L-SN-V2 + SA + SE		
	S-F-C-Bi-L-SN-V2 + CA + SE		
	S-F-C-Bi-L-SN-V2 + SA + CA + SE		

https://doi.org/10.1371/journal.pone.0348906.t006

Python. The system can use audio and camera to realize emotion recognition. Due to the limitation of conditions, the indoor experiment chooses to use audio and camera monitoring for unsafe emotion recognition test, and the field experiment chooses to use audio and images for unsafe emotion recognition test. All subjects have signed the informed consent form for the experiment.

(1) **Indoor simulation program**

According to the improvement strategy in the previous section, the indoor experimental program was designed as shown in Table 7. Aiming at the biggest possible influencing factors of the emotion recognition model in the actual process: sound intensity and facial expression capture degree, different simulation scenarios are formed by the subject controlling the speaking voice size and the distance of the speaker. The experimental environment is an indoor laboratory of Wuhan University of Science and Technology, and the experiment is carried out in an environment with no construction noise, people walking around and other disturbances, and the sound intensity is controlled by the YSD130 noise detector. To ensure experimental fairness, all tests used the proposed model (S-F-C-Bi-L-SN-V2 + SA + CA + SE) with identical network backbone, pre-trained weights, data preprocessing, hyperparameters, and test environment. The sole variable was the input modality, with the unimodal speech test enabling only the speech branch with masked facial input, the unimodal facial test enabling only the facial branch with masked speech input, and the bimodal test activating both branches simultaneously.

Table 7. Indoor experimental program under different sound intensity and distance conditions.

	Model	Sound intensity (dB) + Distance (m)
1	S-C-Bi-L + SA + CA	50 + 0.5
2		50 + 1.0
3		60 + 0.5
4		60 + 1.0
5	F-SN-V2 + SE	50 + 0.5
6		50 + 1.0
7		60 + 0.5
8		60 + 1.0
9	S-F-C-Bi-L-SN-V2 + SA + SE + CA	50 + 0.5
10		50 + 1.0
11		60 + 0.5
12		60 + 1.0

https://doi.org/10.1371/journal.pone.0348906.t007

(2) **Field experiment program**

We conducted field experiments on the optimal bimodal miners' unsafe emotion recognition model selected from numerical and indoor experiments to verify the model's real-world emotion recognition effect. Three experimental subjects covering different attribute information are selected, and the participant information is shown in Table 8. The voice and expression data were captured and recorded by camera, and then processed by audio editing software to obtain the measured data set for emotion recognition, and finally analyzed using the established bimodal miner unsafe emotion recognition model.

3 Results

3.1 Results of numerical experiments

Figs 2 and 3 illustrate the accuracy and loss curves for the training and validation sets from a representative run. As the number of epochs increased, the accuracy of both sets improved consistently while the loss values decreased. The early stopping mechanism was triggered at the 78th epoch after the validation loss remained stable for 10 consecutive iterations, indicating optimal model fitting without unnecessary computation.

Table 9 presents the ablation results of different modules. Specifically, we employed the baseline network (S-F-C-Bi-L-SN-V2) in the experiments, adding frequency masking, coordinate attention and squeeze-excitation mechanisms to the baseline architecture sequentially. It can be observed that the proposed model (S-F-C-Bi-L-SN-V2 + SA + CA + SE) achieves the best performance, with a mean accuracy of 85.56 ± 0.28% and a 95% confidence interval [85.21%, 85.91%].

Table 8. Personal information of experimental participants.

	Height (cm)	Weight (kg)	Age	Marital status	Level of education	Job type	Seniority
A	170	65	25	Unmarried	Junior college	Mechanical electrician	3
B	173	70	36	Married	Junior high school	Mining workers	8
C	175	75	45	Married	Junior high school	Transporter	12

https://doi.org/10.1371/journal.pone.0348906.t008

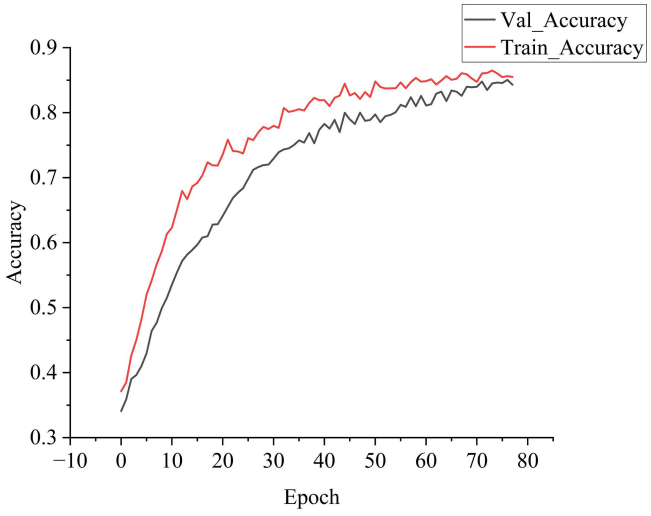

Fig 2. Comparison of Accuracy on Training and Validation.

https://doi.org/10.1371/journal.pone.0348906.g002

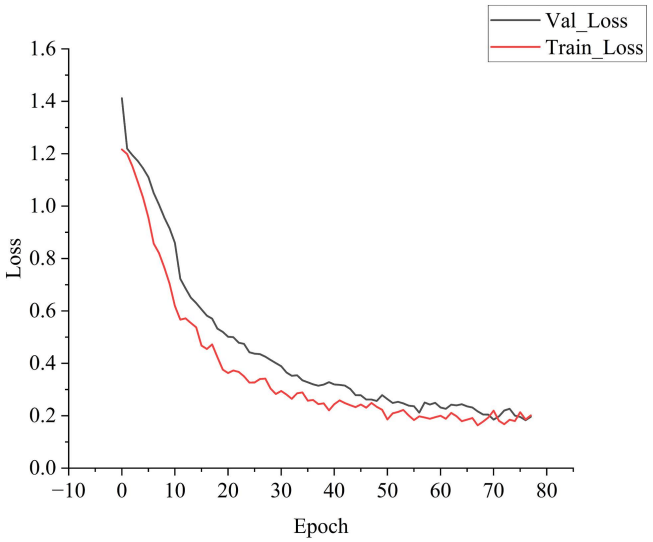

Fig 3. Comparison of Loss Rates on Training and Validation Sets.

https://doi.org/10.1371/journal.pone.0348906.g003

Table 9. Ablation experimental results of different modules.

No	Model	Acc(%)	F1(%)	Loss	95% CI(%)
1	S-F-C-Bi-L-SN-V2	79.36 ± 0.20	78.19 ± 0.19	0.377 ± 0.002	[79.11%, 79.61%]
2	S-F-C-Bi-L-SN-V2 + CA	82.95 ± 0.15	82.71 ± 0.16	0.248 ± 0.002	[82.76%, 83.14%]
3	S-F-C-Bi-L-SN-V2 + SA	82.18 ± 0.17	81.56 ± 0.17	0.270 ± 0.002	[81.97%, 82.39%]
4	S-F-C-Bi-L-SN-V2 + SE	80.89 ± 0.17	81.15 ± 0.18	0.305 ± 0.002	[80.68%, 81.10%]
5	S-F-C-Bi-L-SN-V2 + SA + CA	85.02 ± 0.16	84.88 ± 0.16	0.238 ± 0.002	[84.82%, 85.22%]
6	S-F-C-Bi-L-SN-V2 + SA + SE	84.41 ± 0.17	84.12 ± 0.17	0.256 ± 0.002	[84.20%, 84.62%]
7	S-F-C-Bi-L-SN-V2 + CA + SE	83.10 ± 0.17	82.79 ± 0.17	0.270 ± 0.002	[82.89%, 83.31%]
8	S-F-C-Bi-L-SN-V2 + SA + CA + SE	85.56 ± 0.28	85.20 ± 0.27	0.223 ± 0.003	[85.21%, 85.91%]

https://doi.org/10.1371/journal.pone.0348906.t009

Compared with the baseline (79.36 ± 0.20%), the full model improves accuracy by 6.20 percentage points. All improved models also outperform the baseline, with narrow confidence intervals indicating stable performance across different data splits.

To assess the statistical significance of these improvements, we performed paired t-tests based on the five independent runs using accuracy as the evaluation metric. As summarized in Table 10, every improved model shows a statistically significant gain over the baseline (all $p < 0.001$). Furthermore, the proposed model also achieves a significant improvement over the second-best model (S-F-C-Bi-L-SN-V2 + SA + CA), with a mean difference of 0.54 percentage points ($t = 2.93$, $p = 0.043$). These results confirm that each proposed enhancement strategy contributes meaningfully to recognition accuracy, and that the integrated model provides the best overall performance with statistical reliability.

3.2 Results of indoor experiments

We conducted indoor experiments according to the program in Table 7. Fig 4 presents the system interface during real-time recognition. The model demonstrated high efficiency, with an average inference latency of approximately 0.20s per frame under various environmental configurations. It should be noted that all facial images presented in this manuscript

Table 10. Statistical significance analysis based on Accuracy.

Comparison	Mean Difference (%)	t-value	p-value
+CA vs Baseline	3.59	22.2	<0.001
+SA vs Baseline	2.82	16.5	<0.001
+SE vs Baseline	1.53	11.8	<0.001
+SA+CA vs Baseline	5.66	34.1	<0.001
+SA+SE vs Baseline	5.05	29.7	<0.001
+CA+SE vs Baseline	3.74	22.9	<0.001
+SA+CA+SE vs Baseline	6.20	30.95	<0.001
+SA+CA+SE vs Second-best(+SA+CA)	0.54	2.93	0.043

https://doi.org/10.1371/journal.pone.0348906.t010

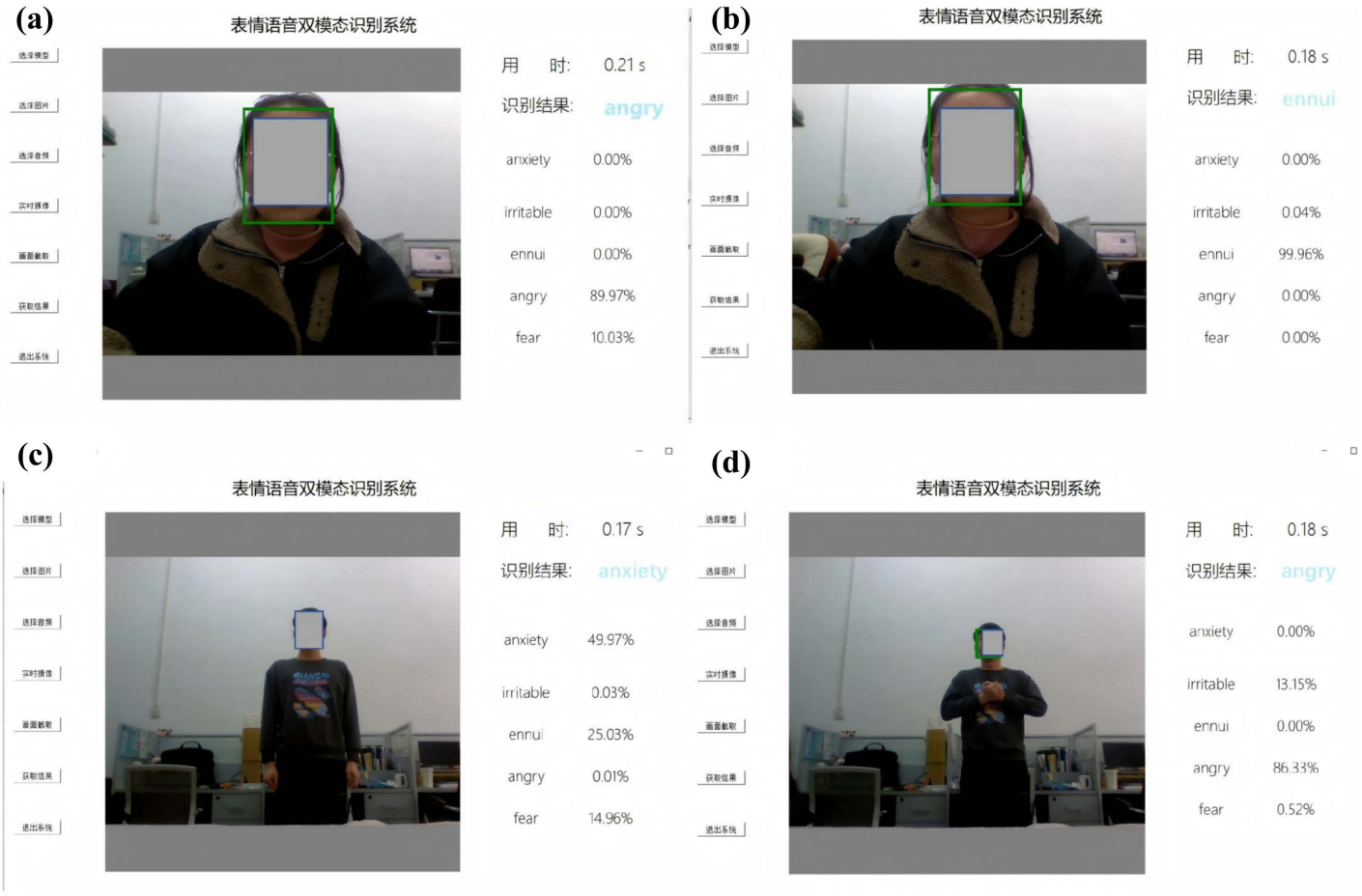

Fig 4. Emotion Detection Recognition Results: (a) 50 dB+0.5 m; (b) 60 dB+0.5 m; (c) 50 dB+1.0 m; (d) 60 dB+1.0 m (Facial regions are blurred for privacy protection; the original unblurred images were used for model training and evaluation).

https://doi.org/10.1371/journal.pone.0348906.g004

have been blurred to protect the privacy of the participants. However, the original unblurred images were used during the training and evaluation stages of the experiments. Therefore, the blurring of images in the manuscript does not affect the reproducibility or validity of the experimental results.

The results in Table 11 show that the improved bimodal miner's unsafe emotion recognition model has the best experimental performance, with the highest recognition accuracy at a sound intensity of 60 dB and a distance of 0.5 m. Compared to unimodal speech and facial expression models in the same scenario, the bimodal approach improved recognition accuracy by 6.38% and 7.87%, respectively. For the simulation scene (sound intensity of 50dB, distance of 1.0m), which has the lowest accuracy among the measured scenes of the bimodal model, outperforming the unimodal speech and expression models by 4.31% and 4.68%, respectively.

3.3 Results of field experiments

In order to preliminarily explore the practical feasibility of the model in real-world mining environments, we carried out field experiments with the informed consent of all participants. As shown in Fig 5(a), we collected audiovisual data from three miners during their daily work. After processing, we obtained speech signals and facial expression fragments with obvious emotional characteristics, and used the optimal model to recognize the emotional state of the three miners.

The average recognition accuracy and distribution of five types of unsafe emotions among the three miners are shown in Fig 5(b) and Fig 5(c)–Fig 5(e). The highest recognition accuracy for the anger emotion category was 81.25%, followed by anxiety and irritability with 76.56% and 73.23%, respectively. Ennui and fear had the lowest recognition accuracies of 70.67% and 65.09%, respectively.

4 Discussion

4.1 Optimally improved model

The experimental results demonstrate that the overall recognition performance of the improved network is consistently superior to that of the baseline model across multiple random data splits. In particular, the proposed model (S-F-C-Bi-L-SN-V2 + SA + CA + SE) achieves the best performance, with an average accuracy of 85.56 ± 0.28% and an F1-score of 85.20 ± 0.27%. Moreover, the relatively small standard deviation and narrow confidence interval demonstrate that the proposed model maintains stable performance under different data partitioning conditions. A comparative analysis of the eight models shows that all three feature enhancement strategies contribute to performance improvement to varying

Table 11. Results for different sound intensity and distance conditions.

	Model	Sound intensity (dB) + distance (m)	Accuracy
1	S-C-Bi-L + SA + CA	50 + 0.5	82.15
2		50 + 1.0	81.78
3		60 + 0.5	85.69
4		60 + 1.0	83.42
5	F-SN-V2 + SE	50 + 0.5	82.95
6		50 + 1.0	81.41
7		60 + 0.5	84.20
8		60 + 1.0	80.05
9	S-F-C-Bi-L-SN-V2 + SA + SE + CA	50 + 0.5	88.45
10		50 + 1.0	86.09
11		60 + 0.5	92.07
12		60 + 1.0	89.13

https://doi.org/10.1371/journal.pone.0348906.t011

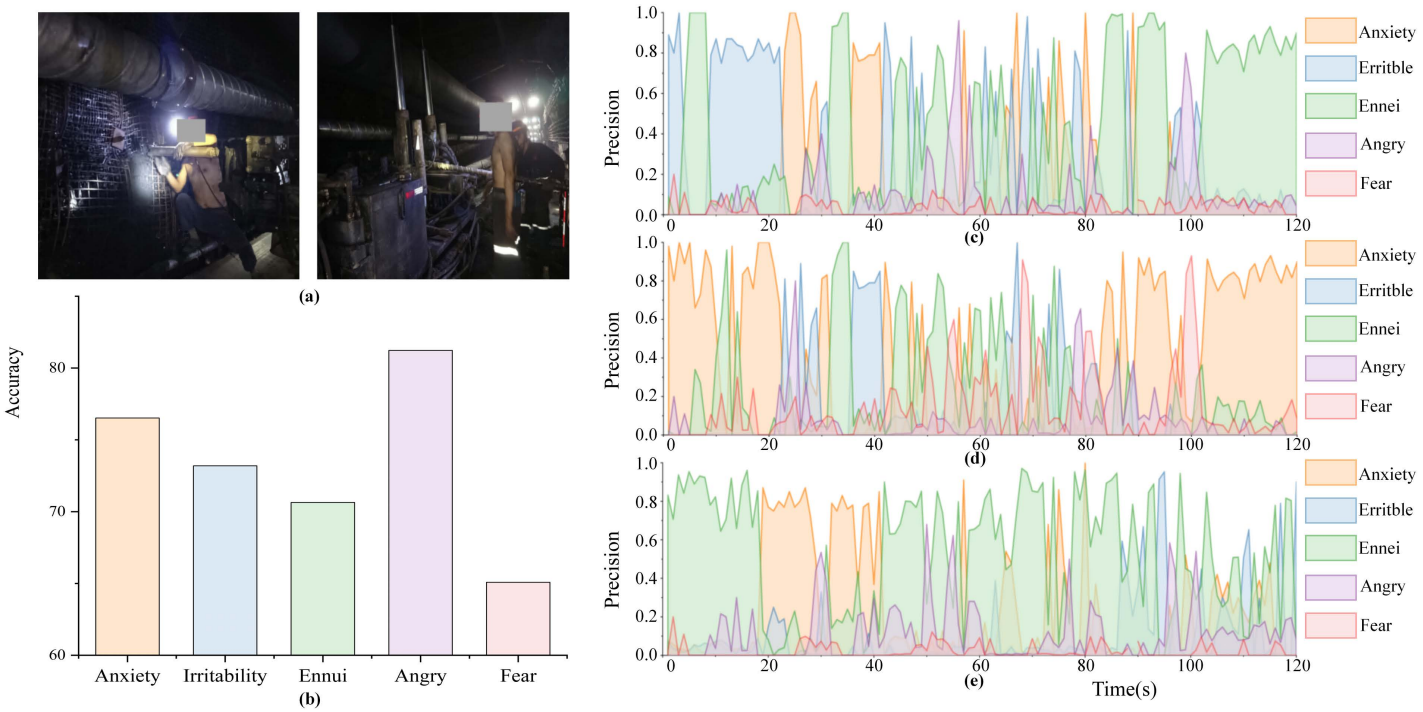

Fig 5. Field experiment results: (a) miner's work site; (b) average recognition accuracy; (c) temporal distribution of unsafe emotions for miner A; (d) miner B; (e) miner C. The individuals shown in this figure have given their written permission to publish their photographs under the CC BY 4.0 license.

https://doi.org/10.1371/journal.pone.0348906.g005

degrees. Among them, the coordinate attention mechanism provides the most significant contribution, followed by frequency masking and squeeze-excitation mechanisms. On average, the coordinate attention module improves accuracy by approximately 3.6% compared with the baseline, while frequency masking and squeeze-excitation bring improvements of approximately 2.8% and 1.5%, respectively.

To validate the effectiveness of the proposed method, we compared it with several state-of-the-art approaches. The accuracy results we compared were all reported in these research works. Table 12 demonstrates that on the Fer2013 dataset, the bimodal model achieved an average accuracy of 84.51%, representing a 0.21% improvement in recognition accuracy over the unimodal MiniXception network. Although there remains a certain performance gap compared to Transformer-based architectures, the proposed model achieves a favorable balance between recognition accuracy and computational efficiency, and has the potential to be applied in underground mining edge monitoring scenarios. Transformer-based models typically require higher computational resources, which limits their deployment in underground mining environments with constrained computing power. On the Ravdess dataset, the bimodal model attained an average accuracy of 88.17%, surpassing the state-of-the-art method by 0.27%. It also demonstrated significant performance advantages over other lightweight models.

4.2 Application analysis of the improved model

(1) The effect of sound intensity and distance in practical application of the model

The results in Table 11 show that at a sound intensity of 50 dB, the accuracy decreases by 2.36% for every 0.5 m increase in distance, and at a sound intensity of 60 dB, the accuracy decreases by 2.94% for every 0.5 m increase in distance.

Table 12. Performance comparison of different methods.

Source	Method	Dataset	Acc
[17]	MiniXception	Fer2013	84.3
[31]	Transformer with graph attention	Fer2013	90.4
[39]	CNN and BiLSTM are used by Transfer Learning	RAVDESS	80.08
[40]	BiLSTM-multi-head attention	RAVDESS	82.42
[41]	Lightweight CNN with Uncertainty Learning	RAVDESS	81.9
[42]	3DCNN	RAVDESS	87.9
[43]	VCAN	RAVDESS	75.9
[44]	Lightweight VGGNet	RAVDESS	86.25
Ours	S-F-C-Bi-L-SN-V2 + SE + SA + CA	Fer2013 RAVDESS	84.51 88.17

https://doi.org/10.1371/journal.pone.0348906.t012

Similarly, at a distance of 0.5m, for every 10dB increase in sound intensity, the accuracy rate increases by 3.62%. At a distance of 1.0m, for every 10dB increase in sound intensity, the accuracy rate increases by 3.04%. From the above series of data changes combined with Fig 6, it can be found that the change of sound intensity contributes more to the accuracy of the improved bimodal model for emotion recognition than the change of distance, i.e., the accuracy is more significantly affected by the size of sound intensity. This phenomenon has received limited attention in previous studies, and our next research work can further evaluate how the interaction of audio and visual information affects emotion recognition accuracy under different sound intensity and distance conditions.

(2) **Other areas where improvements can be made in practical application**

From the results in Fig 5(b), the optimal bimodal model achieves the recognition of different categories of unsafe emotions. The low recognition accuracy of ennui and fear, 70.67% and 65.09%, respectively, may overlap with the features of anger emotions, making it difficult to differentiate, and may also be related to external environmental factors such as the noise generated by the operation of the facilities at the work site of the miners and the wearing of helmets, which reduces

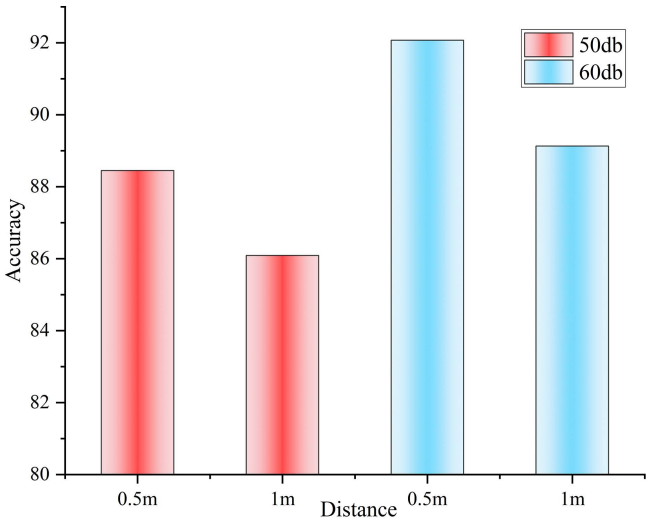

Fig 6. Effect of sound intensity and distance on accuracy.

https://doi.org/10.1371/journal.pone.0348906.g006

the accuracy of the model. Some studies [32–34] have pointed out that the underground environment of coal mines is characterized by dim lighting, dust interference, obstruction, position changes, multiple targets, and other complex factors, which can significantly reduce recognition accuracy. These findings also provide a basis for future improvements of the model.

In addition, although the field data cover typical mining interference factors, due to the relatively small number of participants, the experimental results can be regarded as a preliminary verification of the application potential of the model. Nevertheless, this exploratory field experiment offers a preliminary attempt to evaluate the model under real-world mining conditions. Future research needs to carry out more large-scale field research in more mining areas to further verify the applicability and long-term stability of the model in real-world mining scenarios.

5 Conclusion and future research work

The present study proposes a deep learning-based model for identifying unsafe emotions in miners. In the feature extraction process, three feature enhancement strategies are proposed: First, frequency masking is applied as part of the SpecAugment strategy to enhance speech spectrogram representations. The weight of emotional information is increased through the attention mechanism. The ShuffleNet-V2 network is optimised in combination with the SE mechanism. Subsequently, group comparisons were conducted using numerical experiments. The results of the study show that these three feature enhancement strategies contribute to performance improvements. The proposed model (S-F-C-Bi-L-SN-V2 + SA + CA + SE) achieves the best overall performance, with an average accuracy of 85.56 ± 0.28% and an F1-score of 85.20 ± 0.27%. The narrow confidence interval and low variance indicate that the proposed method exhibits stable performance across different data partitioning conditions.

It is noteworthy that, during real-world testing, it was discovered that sound intensity has a more significant impact on recognition accuracy – a phenomenon that has received limited attention in previous studies. Consequently, these issues merit further optimisation and exploration in subsequent research. Additionally, Future work will focus on expanding the dataset to include large-scale real-world mining data, exploring a more refined emotional classification system, and further validating the applicability and long-term stability of the model through more extensive field experiments. In addition, transformer-based cross-modal fusion strategies will be investigated to further improve robustness under extreme noise and occlusion conditions.

Author contributions

Conceptualization: Ying Lu.

Data curation: Xinyv Shi.

Funding acquisition: Ying Lu.

Methodology: Zihao Zhao.

Project administration: Ying Lu, Liang Yan.

Resources: Liang Yan.

Software: Zihao Zhao, Xinyv Shi.

Supervision: Ying Lu, Liang Yan.

Validation: Zihao Zhao, Xinyv Shi.

Visualization: Xinyv Shi.

Writing – original draft: Zihao Zhao.

Writing – review & editing: Ying Lu.

References

1. Liang Y, Kleijn R, Tukker A, van der Voet E. Material requirements for low-carbon energy technologies: a quantitative review. Renewable and Sustainable Energy Reviews. 2022;161:112334. https://doi.org/10.1016/j.rser.2022.112334
2. Li C, Wang A, Chen X, Chen Q, Zhang Y, Li Y. Regional distribution and sustainable development strategy of mineral resources in China. Chin Geogr Sci. 2013;23(4):470–81. https://doi.org/10.1007/s11769-013-0611-z
3. Sun X, Liu Y, Guo S, Wang Y, Zhang B. Interregional supply chains of Chinese mineral resource requirements. Journal of Cleaner Production. 2021;279:123514. https://doi.org/10.1016/j.jclepro.2020.123514
4. National Bureau of Statistics of China. Statistical bulletin on national economic and social development in 2023. Beijing: National Bureau of Statistics of China; 2024. https://www.stats.gov.cn/xxgk/sjfb/tjgb2020/202402/t20240229_1947923.html
5. Yu M, Qin W, Li J. The influence of psychosocial safety climate on miners' safety behavior: a cross-level research. Safety Science. 2022;150:105719. https://doi.org/10.1016/j.ssci.2022.105719
6. Ali M, Pal I. Assessment of workers' safety behavior in the extractive industries: the case of underground coal mining in Pakistan. The Extractive Industries and Society. 2022;10:101087. https://doi.org/10.1016/j.exis.2022.101087
7. Kong T, Fang W, Love PED, Luo H, Xu S, Li H. Computer vision and long short-term memory: learning to predict unsafe behaviour in construction. Advanced Engineering Informatics. 2021;50:101400. https://doi.org/10.1016/j.aei.2021.101400
8. Guo T, Wang Y, Qiu SR. Mental health risks and unsafe behaviors of coal mine workers: a quantitative study based on cultural circle theory. China Safety Science Journal. 2021;31(S1):129.
9. Guo HX, Lu YF, Li XZ, Hou JJ, Peng K. Optimization research on emergency evacuation of coal mine underground in sub-region based on social force model. Safety and Environmental Engineering. 2023;30(05):121–32.
10. Kim J, André E. Emotion recognition based on physiological changes in music listening. IEEE Trans Pattern Anal Mach Intell. 2008;30(12):2067–83. https://doi.org/10.1109/TPAMI.2008.26 PMID: 18988943
11. Bianco S, Napoletano P. Biometric recognition using multimodal physiological signals. IEEE Access. 2019;7:83581–8. https://doi.org/10.1109/access.2019.2923856
12. Kou L, Tao Y, Kwan M-P, Chai Y. Understanding the relationships among individual-based momentary measured noise, perceived noise, and psychological stress: a geographic ecological momentary assessment (GEMA) approach. Health Place. 2020;64:102285. https://doi.org/10.1016/j.healthplace.2020.102285 PMID: 32819555
13. Akinlade O, Vakaj E, Dridi A, Tiwari S, Ortiz-Rodriguez F. Semantic segmentation of the lung to examine the effect of COVID-19 using UNET model. 2022.
14. Taiwo G, Vadera S, Alameer A. Vision transformers for automated detection of pig interactions in groups. Smart Agricultural Technology. 2025;10:100774. https://doi.org/10.1016/j.atech.2025.100774
15. Ajagbe S, Oladosu J, Olayiwola A, Falohun A. Design and development of automatic speech recognition (ASR) system for low-resource language using convolutional neural network model. Journal of Computer Science and its Applications. 2025;31(2):10–8.
16. He XX. Research on unsafe behavior recognition of underground coal miners based on machine vision. Xi'an: Xi'an University of Architecture and Technology; 2024.
17. Chen Y, Liu Y, Lu C, Chai P, Li S, Zou Y. Research on work-stress recognition for deep ground miners based on depth-separable convolutional neural network. Journal of Loss Prevention in the Process Industries. 2024;91:105410. https://doi.org/10.1016/j.jlp.2024.105410
18. Kang D, Kim D, Kang D, Kim T, Lee B, Kim D, et al. Beyond superficial emotion recognition: modality-adaptive emotion recognition system. Expert Systems with Applications. 2024;235:121097. https://doi.org/10.1016/j.eswa.2023.121097
19. Nidhi, Verma B. From methods to datasets: a detailed study on facial emotion recognition. Appl Intell. 2023;53(24):30219–49. https://doi.org/10.1007/s10489-023-05052-y
20. Huang ZY. Research on miners' mental state recognition based on multi-source information fusion algorithms. Xi'an: Xi'an University of Science and Technology; 2022.
21. Yang H, Chen CLP, Chen B, Zhang T. Improving the interpretability through maximizing mutual information for EEG emotion recognition. IEEE Trans Affective Comput. 2025;16(2):744–57. https://doi.org/10.1109/taffc.2024.3463469
22. Wang S, Zhang X, Zhao R. Lightweight CNN-CBAM-BiLSTM EEG emotion recognition based on multiband DE features. Biomedical Signal Processing and Control. 2025;103:107435. https://doi.org/10.1016/j.bspc.2024.107435
23. Li HW. Mechanism and pre-control of the influence of negative emotions on unsafe behavior of coal mine employees [Doctoral dissertation]. Zhengzhou: North China University of Water Resources and Electric Power; 2023.
24. Shukla J, Barreda-Angeles M, Oliver J, Nandi GC, Puig D. Feature extraction and selection for emotion recognition from electrodermal activity. IEEE Trans Affective Comput. 2021;12(4):857–69. https://doi.org/10.1109/taffc.2019.2901673
25. Pan B, Hirota K, Jia Z, Dai Y. A review of multimodal emotion recognition from datasets, preprocessing, features, and fusion methods. Neurocomputing. 2023;561:126866. https://doi.org/10.1016/j.neucom.2023.126866

26. Mehrish A, Majumder N, Bharadwaj R, Mihalcea R, Poria S. A review of deep learning techniques for speech processing. Information Fusion. 2023;99:101869. https://doi.org/10.1016/j.inffus.2023.101869
27. Khare SK, Blanes-Vidal V, Nadimi ES, Acharya UR. Emotion recognition and artificial intelligence: a systematic review (2014–2023) and research recommendations. Information Fusion. 2024;102:102019. https://doi.org/10.1016/j.inffus.2023.102019
28. Zhang J, Xing L, Tan Z, Wang H, Wang K. Multi-head attention fusion networks for multi-modal speech emotion recognition. Computers & Industrial Engineering. 2022;168:108078. https://doi.org/10.1016/j.cie.2022.108078
29. Xavier SP, John SP. Advanced emotion recognition from facial images and speech signals using complex-value spatio-temporal graph convolutional neural networks. SIViP. 2025;19(11). https://doi.org/10.1007/s11760-025-04449-1
30. Li Y-K, Meng Q-H, Wang Y-X, Hou H-R. MMFN: emotion recognition by fusing touch gesture and facial expression information. Expert Systems with Applications. 2023;228:120469. https://doi.org/10.1016/j.eswa.2023.120469
31. Yousafzai SN, Nasir IM, Saidani O, Ghodhbani R, Gu Y, Syafrudin M, et al. A multi-scale simplicial transformer with graph attention for facial emotion recognition. Ain Shams Engineering Journal. 2025;16(10):103584. https://doi.org/10.1016/j.asej.2025.103584
32. Wang P, Zhao HJ. Visual navigation technology based on RGBD in coal mine scene. Safety in Coal Mines. 2022;53(11):136–40.
33. Lu JF, Yang CY. Research on identifying illegal behavior of miners riding overhead passenger devices based on M3CFC-YOLOv7-tiny. Safety in Coal Mines. 2024;55(11):250–6.
34. Wang JF, Duan SY, Pan HG, Jing NB. Lightweight pose estimation spatial-temporal enhanced graph convolutional model for miner behavior recognition. Journal of Mine Automation. 2024;50(11):34–42.
35. Davis S, Mermelstein P. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Trans Acoust, Speech, Signal Process. 1980;28(4):357–66. https://doi.org/10.1109/tassp.1980.1163420
36. Wang CL, Wu CL, Li CW, Zhu MF. Foreign object recognition based on coordinated attention and atrous convolution. Computer Systems & Applications. 2024;33(03):178–86.
37. Hou Q, Zhou D, Feng J. Coordinate attention for efficient mobile network design. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021. p. 13708–17. https://doi.org/10.1109/cvpr46437.2021.01350
38. Geetha A, Mala T, Priyanka D, Uma E. Multimodal emotion recognition with deep learning: advancements, challenges, and future directions. Information Fusion. 2024;105:102218. https://doi.org/10.1016/j.inffus.2023.102218
39. Luna-Jiménez C, Griol D, Callejas Z, Kleinlein R, Montero JM, Fernández-Martínez F. Multimodal emotion recognition on RAVDESS dataset using transfer learning. Sensors (Basel). 2021;21(22):7665. https://doi.org/10.3390/s21227665 PMID: 34833739
40. Jin Z, Zai W. Audiovisual emotion recognition based on bi-layer LSTM and multi-head attention mechanism on RAVDESS dataset. J Supercomput. 2024;81(1). https://doi.org/10.1007/s11227-024-06582-z
41. Radoi A, Cioroiu G. Uncertainty-based learning of a lightweight model for multimodal emotion recognition. IEEE Access. 2024;12:120362–74. https://doi.org/10.1109/access.2024.3450674
42. Mocanu B, Tapu R, Zaharia T. Multimodal emotion recognition using cross modal audio-video fusion with attention and deep metric learning. Image and Vision Computing. 2023;133:104676. https://doi.org/10.1016/j.imavis.2023.104676
43. Chen R, Zhou W, Li Y, Zhou H. Video-based cross-modal auxiliary network for multimodal sentiment analysis. IEEE Trans Circuits Syst Video Technol. 2022;32(12):8703–16. https://doi.org/10.1109/tcsvt.2022.3197420
44. Akinpelu S, Viriri S, Adegun A. Lightweight deep learning framework for speech emotion recognition. IEEE Access. 2023;11:77086–98. https://doi.org/10.1109/access.2023.3297269