2025 International Conference on Intelligent Computing
July 26-29, Ningbo, China
https://www.ic-icc.cn/2025/index.php

MAVSA-DI: Mongolian Audio-Visual Sentiment Analysis Based on Deep Residual Shrinkage Network and Improved 3D-DenseNet

Ren Qing-Dao-Er-Ji [1]，Qian Bo[2(✉)]，Ying Lu[3]，Yatu Ji[3] and Nier Wu[3]

[1] School of Information Engineering, Inner Mongolia University of Technology, Hohhot 010000, China
renqingln@imut.edu.cn
[2] School of Information Engineering, Inner Mongolia University of Technology, Hohhot 010000, China
17860363387@163.com
[3] School of Information Engineering, Inner Mongolia University of Technology, Hohhot 010000, China

Abstract. To address the issue of inaccurate extraction of key emotional features in Mongolian audio and video data, which leads to suboptimal sentiment classification performance, this paper proposes a Mongolian Audio-Visual Sentiment Analysis model based on Deep Residual Shrinkage Network and Improved 3D-DenseNet (MAVSA-DI). Specifically, the audio branch adopts a Deep Residual Shrinkage Network (DRSN) to suppress noise interference through a soft-thresholding mechanism and enhance the extraction of emotion-relevant acoustic features. The video branch employs an Improved 3D-DenseNet (I3DD) by integrating the SPD-Conv module, which combines the deep feature extraction capability of SPD-Conv with the dense connectivity of 3D-DenseNet to improve spatiotemporal feature learning from low-resolution facial expressions. Furthermore, Intra-Modal Attention (IMA) mechanisms are applied to both branches to highlight intra-modal key information, followed by Cross-Modal Attention (CMA) to facilitate effective feature fusion. Experimental results demonstrate that the proposed model significantly outperforms several advanced baselines in terms of classification accuracy for Mongolian Audio-Visual Sentiment Analysis (MAVSA).

Keywords: Deep Residual Shrinkage Network, Improved 3D-DenseNet, SPD-Conv, Feature Fusion, Mongolian Audio-Visual Sentiment Analysis.

1 Introduction

With the rapid development of the Internet and artificial intelligence, sentiment analysis has emerged as an increasingly prominent research focus. Although substantial progress has been made in unimodal sentiment analysis based on either audio or visual data, single-modal information often fails to fully and accurately capture the complexity of human emotional states. This limitation becomes particularly pronounced when

dealing with diverse languages, cultures, and expression styles, where accuracy and adaptability remain challenging. In contrast, multimodal sentiment analysis, by leveraging complementary cues from multiple data sources, has demonstrated significant potential in enhancing the accurate interpretation and understanding of affective tendencies [1].

Current research in sentiment analysis predominantly focuses on high-resource languages such as English and Chinese, while studies involving low-resource languages remain relatively scarce. In particular, Mongolian, as an agglutinative language [2], exhibits complex morphological variations and distinctive prosodic characteristics, which substantially increase the difficulty of extracting emotionally relevant representations from both audio and visual modalities. These challenges have significantly limited the applicability and performance of conventional sentiment analysis models in Mongolian language scenarios. To address these issues, this paper proposes a Mongolian Audio-Visual Sentiment Analysis model based on Deep Residual Shrinkage Network and Improved 3D-DenseNet (MAVSA-DI). The major contributions of this work can be summarized as follows:

- A novel audio-visual feature extraction approach is proposed. For the audio modality, a Deep Residual Shrinkage Network (DRSN) is employed to suppress noise through a soft-thresholding mechanism, thereby enhancing the extraction of emotionally salient acoustic features. For the visual modality, an Improved 3D-DenseNet (I3DD) is employed, where an SPD-Conv is embedded to boost the model's capacity for capturing deep and subtle emotional cues from low-resolution facial images.
- An audio-visual feature fusion strategy is proposed. First, Intra-Modal Attention (IMA) mechanisms are applied to enhance salient features within each modality. Then, Cross-Modal Attention (CMA) is employed to facilitate effective integration of audio and visual representations.
- MAVSA-DI is proposed. Experimental results demonstrate that the proposed model significantly enhances the accuracy of Mongolian Audio-Visual Sentiment Analysis (MAVSA).

2 Related Work

MAVSA is a subfield of multimodal sentiment analysis, which aims to integrate information from multiple modalities to achieve more accurate and comprehensive understanding of emotions. Multimodal sentiment analysis has shown great potential in a variety of application domains, including intelligent human-computer interaction [3] and medical diagnosis [4]. At its core, multimodal sentiment analysis relies on two fundamental components: the extraction of unimodal features and their effective fusion.

In terms of audio sentiment analysis, Sun et al. [5] proposed a hybrid framework that combines Complementary Mode-Optimized Empirical Mode Decomposition (CM-OMEMD) with wavelet scattering networks. This approach enhances the extraction of emotion-related acoustic patterns through multi-scale signal decomposition and learning of geometrically invariant features. Li et al. [6] introduced a multi-scale

Transformer for speech emotion recognition, which strengthens the model's ability to learn localized emotional representations across different temporal resolutions.

For visual sentiment analysis, Liang et al. [7] developed a Deep Metric Network with Heterogeneous Semantics (DMN-HS), which innovatively incorporates image captions into the sentiment analysis process to provide a more holistic semantic interpretation of visual content. In addition, Alzamzami et al. [8] proposed a Transformer-based real-time sentiment analysis system that specifically targets the challenges of data heterogeneity and few-shot learning in open-domain social media contexts. Their work provides a comprehensive technical framework for imbalanced and low-resource sentiment analysis.

In the field of multimodal sentiment analysis, Mocanu et al. [9] proposed an end-to-end multimodal emotion recognition framework that focuses on audio-visual fusion. Their method incorporates spatial, channel, and temporal attention mechanisms within a 3D-CNN for visual data and a 2D-CNN for audio data, enabling precise capture of intra-modal features. CMA is then employed to integrate complementary information between the audio and visual modalities. Praveen et al. [10] introduced a novel Incongruity-Aware Cross-Modal Attention (IACA) model, which manages modality mismatches by leveraging complementary relationships. Their approach employs a two-stage gating mechanism to adaptively select semantic features, thereby enhancing modality alignment and mitigating the impact of inconsistent signals.

Regarding Mongolian sentiment analysis, Zhao [11] addressed the challenges of limited Mongolian resources, including the scarcity of labeled corpora and the difficulty of transferring Chinese sentiment analysis models. By constructing a Mongolian-Chinese bilingual knowledge alignment using Chinese corpora and applying cross-lingual sentiment analysis techniques, her work effectively mitigates data scarcity and enriches research in Mongolian sentiment analysis. Yang et al. [12], in response to the insufficient extraction and fusion of multimodal features in Mongolian, proposed a cross-modal hierarchical fusion strategy to enhance the integration of audio-visual features, thereby improving the accuracy and robustness of sentiment classification in Mongolian-language settings.

3 Methods

3.1 Overview

The overall architecture of the proposed MAVSA-DI is illustrated in Fig. 1. The model integrates both audio and visual modalities, and consists of four key components: audio feature extraction, visual feature extraction, feature fusion, and sentiment classification. For audio feature extraction, the input signal is first segmented into short-time frames using Librosa, and transformed into the frequency domain via the Short-Time Fourier Transform (STFT) to capture frequency variations over time. Subsequently, Mel-spectrograms and chroma features are extracted to represent pitch and harmonic content. These time-frequency representations are then passed into a DRSN for hierarchical feature extraction, yielding the audio feature representation F_a. For visual feature extraction, facial regions are localized using OpenCV to detect key facial landmarks, from

which geometric features are derived. The facial images are then processed using an I3DD enhanced with SPD-Conv modules, enabling deep extraction of spatial-temporal emotional cues, particularly from low-resolution facial inputs. This process generates the visual feature representation F_v. Next, IMA mechanisms are applied separately to F_a and F_v to emphasize modality-specific salient information. Following this, a CMA mechanism is employed to enable interaction and fusion between the audio and visual modalities, resulting in the final fused representation F_{fusion}. Finally, the fused features F_{fusion} are passed through a fully connected layer followed by a $Softmax$ function to perform sentiment classification and output the predicted emotion label.

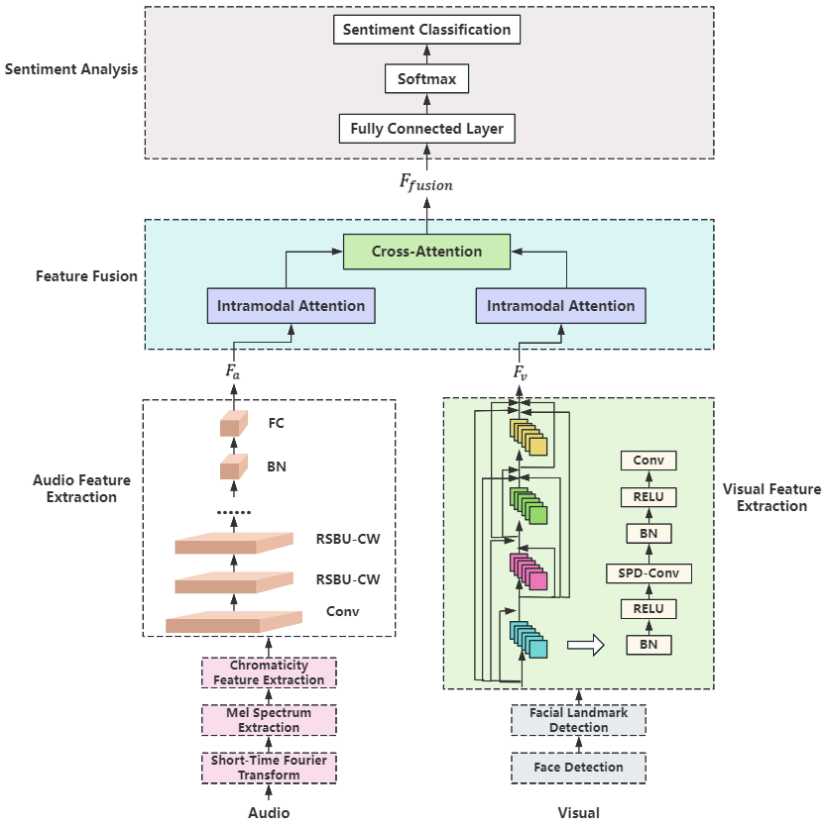

Fig. 1. The overall structure of MAVSA-DI model.

3.2 Audio Feature Extraction

First, the original Mongolian audio data are preprocessed using Librosa. During this process, the audio signals are initially transformed from the time domain to the frequency domain via the STFT, which enables the preliminary capture of time-varying frequency patterns and provides a clear representation of the spectral composition at

each time frame—thus laying the foundation for subsequent feature extraction. Next, the frequency-domain signals are mapped onto the Mel frequency scale to better align with human auditory perception, and chroma features are further extracted to more accurately reflect the pitch, prosody, and their temporal variations in Mongolian speech.

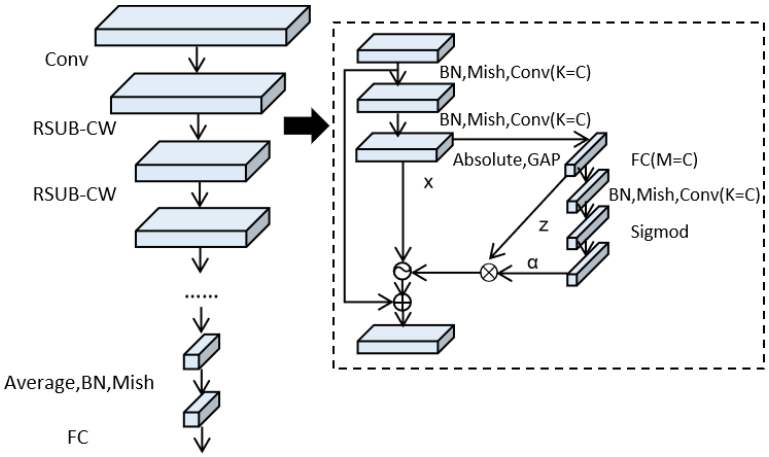

Fig. 2. Deep Residual Shrinkage Network.

Subsequently, the extracted time–frequency features are fed into a DRSN for deep-level feature modeling. As illustrated in Fig. 2, the DRSN consists of multiple convolutional layers, each equipped with a set of kernels designed to capture local patterns in the audio signal.

The output of the $l-th$ convolutional layer can be formulated as:

$$Y^l = \sum_{i=1}^{n^l} K_i^l * X^l \tag{1}$$

Where X^l, K_i^l, and Y^l denote the input, convolution kernel, and output of the $l-th$ convolutional layer, respectively. n^l represents the number of convolution kernels in the $l-th$ layer, and the $*$ indicates the convolution operation.

During the training of deep neural networks, issues such as vanishing or exploding gradients often arise. To address this, the DRSN introduces a residual connection mechanism, in which the input of a previous layer is added directly to the output of a later layer, bypassing intermediate layers. This design effectively mitigates gradient degradation and enables the construction of deeper architectures for capturing complex emotional patterns in audio data. The output of the $m-th$ residual block can be expressed as:

$$C^m = A^m + B^m \tag{2}$$

Where A^m denotes the input to the $m-th$ residual block, and B^m represents the output obtained after processing through its intermediate layers.

In practice, raw audio data often contain a significant amount of noise and redundant information that is irrelevant to sentiment analysis, which can interfere with model

training and reduce classification accuracy. To address this, the DRSN incorporates a shrinkage mechanism within its residual blocks. Specifically, it employs techniques such as soft thresholding to suppress low-importance or irrelevant features while emphasizing emotionally salient information. This design enhances both the robustness and generalization capability of the model. Finally, a pooling layer is applied to downsample the extracted features, further improving computational efficiency and yielding high-quality audio representations for subsequent sentiment classification.

3.3 Visual Feature Extraction

OpenCV is employed to extract low-level visual features, laying the groundwork for subsequent facial emotion representation. During the face detection stage, Haar cascade classifiers provided by OpenCV are used to accurately locate facial regions. Subsequently, facial landmark detection is applied to identify key points within the detected face areas. By computing geometric parameters such as the distances and angles between these landmarks, a set of facial geometric features is derived.

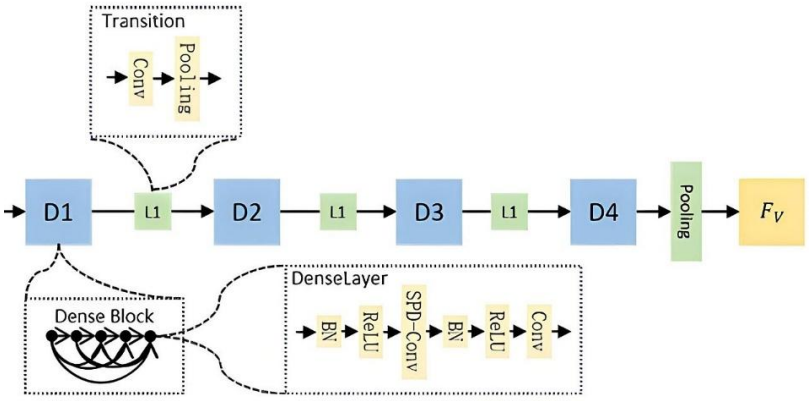

Fig. 3. Improved 3D-DenseNet structure diagram.

Subsequently, the cropped facial images are processed using an I3DD.This improved version integrates an SPD-Conv module, which enhances the model's ability to capture spatial–temporal features of facial expressions, particularly in low-resolution scenarios, thereby addressing limitations of traditional models in handling facial imagery. As illustrated in Fig. 3, the input image passes through multiple dense blocks. Each dense block consists of several stacked DenseLayers, where each layer is densely connected to all preceding layers. The input data are first normalized using Batch Normalization (BN) to achieve zero mean and unit variance, followed by learnable affine transformations to scale and shift the normalized data. Then, a Rectified Linear Unit (ReLU) activation function is applied to introduce non-linearity by zeroing out negative values and retaining positive ones.

The SPD-Conv is capable of capturing the symmetric positive definite structure embedded in data, offering more precise representation of deep features compared to

conventional convolution operations. In the context of complex facial image data, SPD-Conv exhibits a stronger capacity to capture subtle variations in facial expressions and to extract underlying structural information. Let K denote the convolutional kernel of the layer; the convolution operation can be formulated as:

$$(F * K)(i,j) = \sum_m \sum_n F(i+m, j+n) K(m,n) \tag{3}$$

Where F denotes the input feature map, and (i,j) represents the spatial index of the output feature map. Through this convolution operation, the kernel K slides over the input feature map and performs a weighted summation at each position, thereby enabling the extraction of more informative and discriminative features.

Transition modules are employed between dense blocks to maintain consistency in feature map dimensionality. Upon completion of all dense block operations, the network outputs the visual sentiment feature F_v, which encapsulates rich emotional information embedded in facial images and serves as a strong foundation for subsequent sentiment classification tasks.

By integrating the SPD-Conv into specific layers of the 3D-DenseNet architecture, the model effectively combines the complementary strengths of both components. The powerful deep feature extraction capability of SPD-Conv, when combined with the densely connected structure of 3D-DenseNet, significantly enhances the precision and accuracy of spatiotemporal feature learning. This integration not only enriches the model's feature representation but also substantially improves its performance in sentiment classification tasks, enabling more accurate recognition of facial emotions under complex real-world conditions.

3.4 Feature Fusion

In the feature fusion stage, an IMA mechanism is introduced to emphasize salient information within both the audio and visual features.

For the audio modality, the audio features F_a extracted by the DRSN are first linearly transformed to generate the query Q_a, key K_a, and value V_a vectors. The attention weights α_a in the audio modality reflect the relative importance of each feature in contributing to emotion representation. The computation is defined as follows:

$$\alpha_a = Softmax\left(\frac{Q_a K_a^T}{\sqrt{d_k}}\right) \tag{4}$$

Where d_k denotes the dimensionality of the key vectors. Subsequently, the enhanced audio feature is obtained through a weighted summation as follows:

$$F_a' = \alpha_a V_a \tag{5}$$

This mechanism enables the model to more effectively focus on the segments of the audio that are highly relevant to emotional expression, thereby enhancing the representation quality of the audio features.

In the visual modality, a similar approach is applied to enhance the attention on the visual features F_v extracted by the I3DD. Query vectors Q_v, key vectors K_v, and value

vectors V_v are obtained through linear transformations. The attention weights are computed as follows:

$$\alpha_v = Softmax\left(\frac{Q_v K_v^T}{\sqrt{d_k}}\right) \tag{6}$$

The attention-weighted visual features are then obtained accordingly:

$$F_v' = \alpha_v V_v \tag{7}$$

Through this approach, the key information within the visual features is enhanced, which contributes to improving their representational capacity in sentiment analysis.

After completing the IMA enhancement, the enhanced audio features F_a' and visual features F_v' undergo CMA interaction to facilitate information sharing and complementarity between the two modalities. First, linear transformations are applied to F_a' and F_v' to obtain the cross-modal query vectors Q_{av} , key vectors K_{av} , and value vectors V_{av}. Then, the CMA weights are computed as follows:

$$\beta_{av} = Softmax\left(\frac{Q_{av} K_{av}^T}{\sqrt{d_k}}\right) \tag{8}$$

The final audio-visual fused feature F_{fusion} is obtained through a weighted summation.

3.5 Sentiment Analysis

The F_{fusion} is fed into a fully connected layer to produce the raw score vector for emotion categories, $y_{pre} \in R^7$. Subsequently, the $Softmax$ function is applied to normalize y_{pre}, which is defined as follows:

$$Softmax(y_{pre,i}) = \frac{e^{y_{pre,i}}}{\sum_{j-1}^{7} e^{y_{pre,j}}} \tag{9}$$

Here, $y_{pre,i}$ denotes the raw score corresponding to the $i-th$ emotion category.

The final emotion classification result is then obtained as follows:

$$y_{final} = Softmax(y_{pre}) \tag{10}$$

4 Experiment

4.1 Dataset

The dataset used in this study is a Mongolian multimodal emotion dataset recorded by the Artificial Intelligence Laboratory of Inner Mongolia University of Technology. It contains seven emotion categories: sad, angry, surprise, fear, happy, disgusted, and neutral. Each category includes 300 samples, resulting in a total of 2100 video clips.

The audio tracks were extracted from the video dataset and saved as .wav files. The duration of audio samples for each emotion category is summarized in Table 1.

Table 1. Audio duration of different emotional categories.

Emotions	The shortest duration (seconds)	The longest duration (seconds)	Total duration (hours)
Sad	2.90	9.71	0.43
Angry	1.93	5.67	0.24
Surprise	1.88	5.50	0.26
Fear	1.84	8.52	0.29
Happy	2.46	9.52	0.37
Disgusted	2.02	5.85	0.28
Neutral	2.90	8.17	0.37

The videos were processed into frame sequences through frame-by-frame extraction. Due to variations in video duration and content complexity across different emotion categories, the number of extracted frames differs accordingly. The number of frames extracted for each emotion category is presented in Table 2.

Table 2. The number of framed images of different emotional categories.

Emotions	Number of framed images (sheet)
Sad	23359
Angry	7977
Surprise	9111
Fear	13402
Happy	15652
Disgusted	10305
Neutral	16173

4.2 Experimental Environment and Evaluation Index

Experimental Environment. As shown in Table 3:

Table 3. Experimental environment configuration.

Experimental environment	Related configuration
Operating System	Ubuntu 18.04.5 LTS
GPU	Nvidia Tesla P100
CPU	Intel® Xeon® Gold 6310 CPU @ 2.10GHz
Memory	64G
Python	3.9
Optimizer	Adam
CUDA	11.2

Evaluation Index. To evaluate the performance of the proposed MAVSA-DI model, this study adopts Accuracy, Recall, and F1-score as evaluation metrics.

4.3 MAVSA-DI Model Experiment

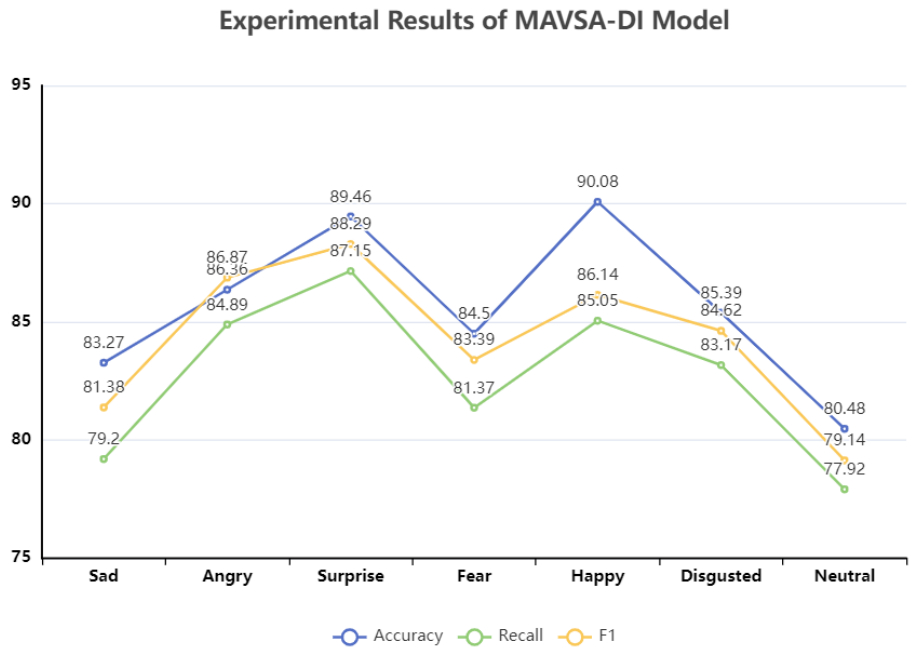

Fig. 4. Experimental results of MAVSA-DI model.

The experimental results presented in Fig. 4 indicate that the MAVSA-DI model exhibits varying recognition performance across different emotion categories. Specifically, the model achieves higher accuracy in recognizing emotions such as 'happy' and 'surprise', suggesting its effectiveness in identifying emotions with distinct expressive cues. In contrast, the accuracy for the 'neutral' emotion is the lowest, which can be attributed to the lack of strong emotional signals, making it more challenging for the model to distinguish relevant features. The performance on other emotion categories remains relatively balanced, demonstrating the model's ability to effectively learn and classify emotions with clear affective tendencies. Overall, the MAVSA-DI model shows promising sentiment recognition capabilities.

4.4 Contrast Experiment

To validate the effectiveness of the proposed MAVSA-DI model, several representative sentiment analysis models were selected for comparative experiments. The baseline models include RNN [13], 3D-DenseNet [14], VCAN [15], and CMAVF-AD [9].

Table 4. The experimental results of different models on the constructed dataset.

Model	Modality	Accuracy	Recall	F1
RNN	Audio	67.24	66.07	66.52
3D-DenseNet	Visual	74.13	73.70	72.50
VCAN	Audio+ Visual	77.10	76.02	76.54
CMAVF-AD	Audio+ Visual	79.42	77.28	78.34
MAVSA-DI (ours)	**Audio+ Visual**	**85.47**	**82.06**	**83.24**

Table 4 shows that the MAVSA-DI model achieves superior performance compared to all baseline models in terms of accuracy (85.47%), recall (82.06%), and F1 (83.24%), highlighting its effectiveness in MAVSA tasks.

In contrast, the traditional RNN model performs the worst under the audio modality, indicating its limited ability to capture emotional features from speech signals. The 3D-DenseNet model yields better results in the visual modality than RNN but still performs significantly worse than models utilizing multimodal fusion. Although both VCAN and CMAVF-AD benefit from integrating audio and visual modalities and exhibit some performance improvement, they fail to sufficiently extract the critical emotional features across modalities. In comparison, MAVSA-DI demonstrates enhanced representational capacity by accurately capturing salient emotional cues, thereby significantly improving classification performance and validating the effectiveness of the proposed model in MAVSA.

4.5 Ablation Experiment

Table 5. Performance comparison of ablation experiments.

Model	Accuracy	Recall	F1
w/o DRSN	71.82	71.06	70.36
w/o I3DD	66.05	63.38	62.74
w/o IMA	76.37	73.84	74.65
w/o CMA	80.29	78.26	79.57
MAVSA-DI (ours)	**85.47**	**82.06**	**83.24**

As shown in Table 5, an ablation study was conducted to evaluate the contribution of each component within the proposed model. The experimental results demonstrate that the complete model (MAVSA-DI) achieves the best overall performance across all evaluation metrics.

When the DRSN module in the audio branch is removed, the model's performance drops significantly, with the F1 decreasing to 70.36%. This indicates that DRSN plays a crucial role in suppressing redundant features and enhancing emotionally salient audio representations. The exclusion of the I3DD module in the visual branch leads to the most substantial performance degradation, with the accuracy reduced to 66.05% and

the F1 to 62.74%, confirming the module's effectiveness in capturing spatiotemporal facial cues, especially in low-resolution settings. Additionally, removing either the IMA mechanism or the CMA mechanism results in noticeable declines in performance, which further validates the importance of attention mechanisms in enhancing modality-specific representations and enabling effective cross-modal interaction. Overall, each component positively contributes to the model's performance, verifying the effectiveness and necessity of the design choices in MAVSA-DI.

5 Conclusion

This study proposes a Mongolian Audio-Visual Sentiment Analysis model based on Deep Residual Shrinkage Network and Improved 3D-DenseNet (MAVSA-DI), which effectively enhances the accuracy of MAVSA. To address the challenge of inaccurate extraction of key emotional features in Mongolian audio-visual data, the model employs DRSN and I3DD to perform deep modeling of audio and visual features, respectively, enabling precise emotional representation. Furthermore, IMA mechanisms are introduced to emphasize salient emotional cues within each modality, and a CMA mechanism is used to achieve effective fusion of audio and visual information for final sentiment classification. In future work, we plan to explore cross-lingual model transfer strategies and incorporate a conflict-aware feature disentanglement module to mitigate inter-modal interference and enhance the complementarity of modalities, particularly for emotion categories with low recognition rates.

Acknowledgments. This work was supported by the National Natural Science Foundation of China (62466044), the Basic Research Business Fee Project of Universities Directly under the Autonomous Region (JY20240062), (ZTY2024072), the 'Youth Science and Technology Talent Support Program' Project of Universities in Inner Mongolia Autonomous Region (NJYT23059) and the Inner Mongolia Natural Science Foundation General Project (2022MS06013).

References

1. Joseph, J., Aneesh, R., Zacharias, J.: Deep learning based emotion recognition in human-robot interaction with multi-modal data. In: AIP Conference Proceedings. vol. 3122. AIP Publishing (2024)
2. Hou, H., Sun, S., Wu, N.: A review of mongolian-chinese neural machine translation research. Computer Science **49**(1), 31–40 (2022)
3. Chen, J., Hu, Y., Lai, Q., Wang, W., Chen, J., Liu, H., Srivastava, G., Bashir, A.K., Hu, X.: Iifdd: Intra and inter-modal fusion for depression detection with multi-modal information from internet of medical things. Information Fusion **102**, 102017 (2024)
4. Abdu, S.A., Yousef, A.H., Salem, A.: Multimodal video sentiment analysis using deep learning approaches, a survey. Information Fusion **76**, 204–226 (2021)
5. Sun, C., Ma, L., Li, H.: Speech emotion recognition based on cm-omemd and wavelet scattering network. Signal Processing **39**(4), 688–697 (2023)
6. Li, Z., Xing, X., Fang, Y., Zhang, W., Fan, H., Xu, X.: Multi-scale temporal transformer for speech emotion recognition. arXiv preprint arXiv:2410.00390 (2024)

7. Liang, Y., Maeda, K., Ogawa, T., Haseyama, M.: Deep metric network via heterogeneous semantics for image sentiment analysis. In: 2021 IEEE International Conference on Image Processing (ICIP). pp. 1039–1043. IEEE (2021)
8. Alzamzami, F., El Saddik, A.: Transformer-based feature fusion approach for multimodal visual sentiment recognition using tweets in the wild. IEEE Access **11**, 47070–47079 (2023)
9. 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 **133**, 104676 (2023)
10. Praveen, R.G., Alam, J.: Incongruity-aware cross-modal attention for audio-visual fusion in dimensional emotion recognition. IEEE Journal of Selected Topics in Signal Processing (2024)
11. Zhao, M.: Mongolian sentiment analysis based on multi-lingual feature sharing and deep convolutional neural network. Master's thesis, Inner Mongolia University of Technology (2024)
12. Yang, Y., He, R.F., et al.: Multi-modal sentiment analysis of mongolian language based on pre-trained models and high-resolution networks. In: 2024 International Conference on Asian Language Processing (IALP). pp. 291–296. IEEE (2024)
13. Zaremba, W., Sutskever, I., Vinyals, O.: Recurrent neural network regularization. arXiv preprint arXiv:1409.2329 (2014)
14. Zhang, C., Li, G., Du, S., Tan, W., Gao, F.: Three-dimensional densely connected convolutional network for hyperspectral remote sensing image classification. Journal of Applied Remote Sensing **13**(1), 016519–016519 (2019)
15. Chen, R., Zhou, W., Li, Y., Zhou, H.: Video-based cross-modal auxiliary network for multimodal sentiment analysis. IEEE Transactions on Circuits and Systems for Video Technology **32**(12), 8703–8716 (2022)