SIN-CHN SCIENTIFIC PRESS PTE.LTD.

Molecular & Cellular Biomechanics 2024, 21(3), 717.
https://doi.org/10.62617/mcb717

Article

Analysis and identification of nocturnal groaning syndrome based on multimodal data

Xiaohui Xu[1,†], Min Yu[2,†], Qing Wang[3,*], Xuemei Gao[2,*], Wenai Song[1], Xu Gong[2], Yi Lei[3]

[1] North University of China, Shanxi 030000, China
[2] Department of Orthodontics, Peking University School and Hospital of Stomatology, Beijing 100081, China
[3] Pharmacovigilance Research Center for information technology and Data Science, Cross-strait Tsinghua Research Institute, Xiamen 361000, China
*** Corresponding author:** Qing Wang, 13641213301@139.com; Xuemei Gao, xmgao@263.net
[†] Xiaohui Xu and Min Yu contributed equally to this work

CITATION

Xu X, Yu M, Wang Q, et al. Analysis and identification of nocturnal groaning syndrome based on multimodal data. Molecular & Cellular Biomechanics. 2024; 21(3): 717.
https://doi.org/10.62617/mcb717

ARTICLE INFO

Received: 5 November 2024
Accepted: 11 November 2024
Available online: 25 November 2024

Abstract: Nocturnal groaning syndrome is a common sleep disorder characterized by irregular groaning or vocalizations during nighttime sleep, representing a significant area of research in sleep disorders. Nocturnal groaning syndrome is a common sleep disorder characterized by irregular groaning or vocalizations during nighttime sleep, representing a significant area of research in sleep disorders. proposes a multimodal recognition approach based on speech, image, and text modalities. The study analyzes audio features using Mel Frequency Cepstral Coefficients (MFCC), which is the most common method for identifying nocturnal groaning syndrome. Coefficients (MFCC), extracts image features with pretrained MobileNetV2, and identifies key physiological signals from text using TF-IDF algorithm. Subsequently, Multimodal Compact Bilinear Pooling (MCB) is employed to fuse audio and image features, and a Text-Image CNN is used to combine image and text features. Support Vector Machine (SVM) is then used to classify the fused multimodal features, and decision-level fusion is performed using weighting criteria. Experimental results demonstrate an identification accuracy of 89.5% on the test set, significantly enhancing the auxiliary diagnostic effectiveness of nocturnurnal diagnosis. Experimental results demonstrate an identification accuracy of 89.5% on the test set, significantly enhancing the auxiliary diagnostic effectiveness of nocturnal groaning syndrome.

Keywords: multimodal fusion; feature fusion; night moaning; pattern recognition

1. Introduction

Nocturnal groaning syndrome (NSR) is a rare sleep disorder that usually occurs in the early stages of non-rapid eye movement (NREM) sleep, with a prevalence of 0.17% in Switzerland and 0.8% in France [1,2], and its symptoms are associated with other sleep disorders, such as sleep apnea, which seriously affects the quality of sleep of patients. Existing diagnosis mainly relies on polysomnography (PSG), but its high cost, complexity of operation and laboratory setting affect diagnostic accuracy. In addition, recognition based on single modality data (e.g., audio or EEG) is susceptible to noise interference and insufficient model generalization capability. To this end, this paper proposes a multimodal recognition method integrating audio, video and physiological signals, which improves the accuracy and robustness of recognition through multimodal feature extraction and fusion, and experimentally verifies the effectiveness and high accuracy of the model for non-invasive detection of nocturnal groaning disorder.

2. Methodology

2.1. Data preprocessing

Since both audio and image data are collected from hospitals, they are prone to noise, making noise reduction a necessary first step. Wavelet denoising is an effective method for audio signal processing, especially showing significant advantages in low signal-to-noise ratio (SNR) sleep data analysis. Compared to traditional filtering methods, wavelet denoising can flexibly handle both high-frequency and low-frequency noise, preserving key signal details and edge information, thus enhancing the fidelity of the model's signal [3–5]. Moreover, the multi-scale decomposition characteristic of wavelet transforms allows for separate noise processing across different scales, enabling fine denoising of details in various frequency bands of the signal.

Gaussian denoising is widely used in image processing and offers several advantages. First, Gaussian denoising can smooth out random noise in images, effectively reducing high-frequency noise components and improving the visual quality of images [6,7]. Additionally, the parameters of the Gaussian filter, such as standard deviation, can be flexibly adjusted, allowing users to optimize based on the intensity and distribution of image noise.

2.2. Design of feature extraction method

2.2.1. Speech feature extraction

Speech signal is a time-varying signal whose characteristic parameters are relatively stable in a short period of time, so it is usually divided into frames, the frame length is usually 10 to 30 ms [8]. In this paper, the Hamming window function is used to add windows to the signal, and its expression is:

$$\omega_n = \begin{cases} 0.54 - 0.46\cos\dfrac{2n\pi}{N-1}, 0 \leqslant n \leqslant N-1 \\ 0, other \end{cases} \quad (1)$$

After frame-splitting, feature extraction of speech segments uses Mel Frequency Cepstrum Coefficients (MFCC) and fuses nonlinear attributes and geometric features at the feature layer. Feature fusion is achieved by means of a Deep Restricted Boltzmann Machine (DBM), which is a deep model consisting of a stack of multiple Restricted Boltzmann Machines (RBMs), each of which consists of a visible layer and a hidden layer, with bi-directional connectivity between neurons. Through training, the DBM is able to learn a deep representation of the input features with a loss function of:

$$L(W, a, b) = -\sum_{i=1}^{m} ln\left(P\left(v^{(i)}\right)\right) \quad (2)$$

Finally, the output feature vector is generated by the activation probability of the hidden layer. The training flow is shown in **Figure 1**.

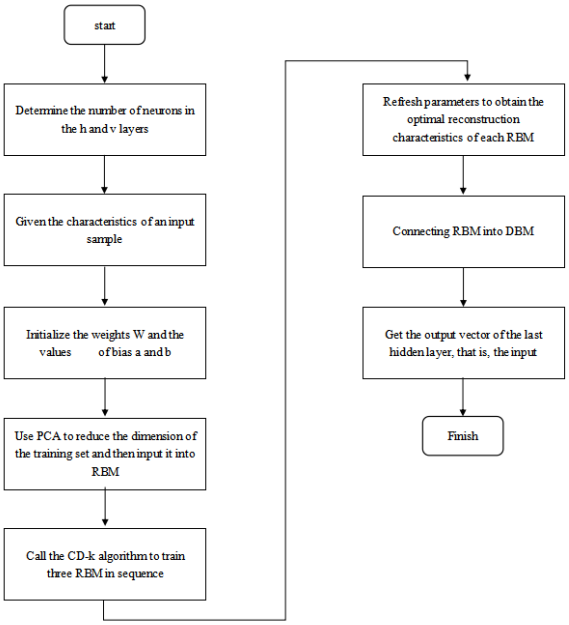

Figure 1. Diagram of DBM training process.

A network consisting of three layers of DBM is built to generate deep fusion features by fusing the selected four types of features. Each layer of DBM consists of three layers of RBM. First, the features are input to the first layer of the DBM for deep fusion and dimensionality reduction to obtain feature 1, feature 2, feature 3, and feature 4 as output from the hidden layer; then, feature 1 is linearly spliced with feature 2 and feature 3 with feature 4, respectively, and input to the second layer of the DBM, which undergoes deep fusion and dimensionality reduction to obtain feature 5 and feature 6; finally, feature 5 and feature 6 go to the third layer of the DBM for further fusion, and finally the deep representation of the input fea tures is generated. The process is shown in **Figure 2**.

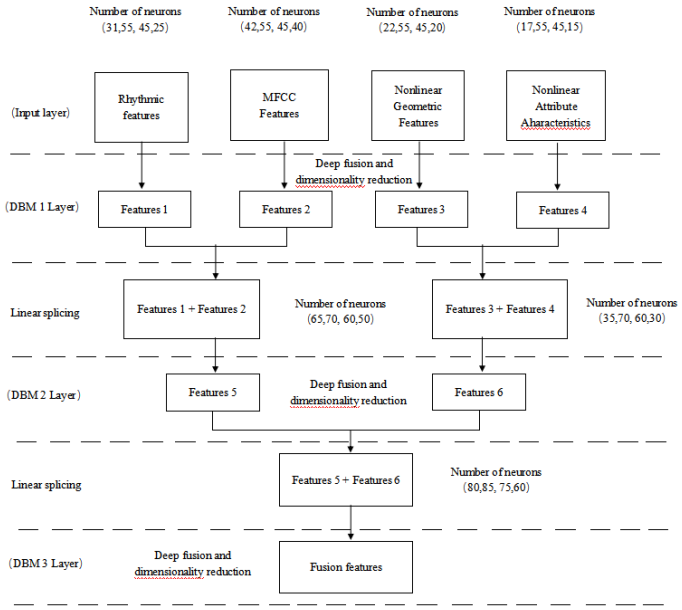

Figure 2. Structure of DBM network.

2.2.2. Image feature extraction

In image feature extraction, we use MobileNetV2 as the main extractor, in order to efficiently obtain meaningful features from complex image data. MobileNetV2 is a lightweight Convolutional Neural Network suitable for mobile devices with efficient computation and low storage requirements [9]. At its core is depth-separable convolution, which divides convolution into deep convolution and point-by-point convolution, significantly reducing the number of parameters and computations while maintaining high classification accuracy [10].

In data preprocessing, the input image is resized to 48 × 48 pixels and normalized to the range of 0 to 1 to improve the stability and speed of training. The main steps of feature extraction include: convolution operation combined with ReLU activation function, batch normalization is used to accelerate convergence, and Global Average Pooling (GAP) [11] reduces the feature map to a single value to provide input to the classification layer. The GAP is computed by the formula:

$$GAP\left(F''\right) = \frac{1}{H \times W} \sum_{h=1}^{H} \sum_{w=1}^{W} F''_{h,w} \tag{3}$$

where H and W denote the height and width of the feature map respectively.The working principle of GAP is shown in **Figure 3**.

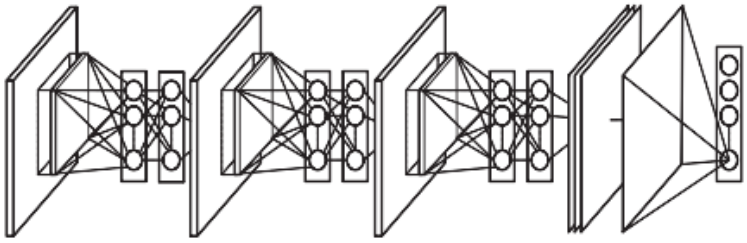

Figure 3. Schematic diagram of the working principle of GAP.

Ultimately, the output of the global average pooling is used as a feature vector and input to the subsequent classifier for efficient image feature extraction and recognition tasks.

2.2.3. Text feature extraction

In a multimodal data fusion framework, text feature extraction is crucial for analyzing physiological signals related to nocturnal groaning disorder. In this paper, the TF-IDF (word frequency-inverse document frequency) algorithm is used to extract text features and measure the importance of a word in a document [12]. TF denotes the frequency of a word's occurrence in a document, defined as the ratio of the number of occurrences of the word to the total number of occurrences of all the words in the document, and IDF measures the importance of a word in the entire set of documents, defined as the logarithm of the ratio of the total number of documents to the number of documents that contain the word. The TF-IDF value is the product of the word frequency and the inverse document frequency, which can effectively extract key textual features from the records of patients with nocturnal groaning disorder and provide support for diagnosis.

2.3. Feature fusion strategy

In the audio-image fusion process, this study employs MCB. Compared to traditional methods such as SVM, MLP, and LSTM networks, MCB demonstrates superior robustness in handling noisy data in real-world scenarios, especially in dealing with potential noise present in hospital-collected audio and image data [13]. MCB effectively reduces the impact of noise during the fusion stage, enhancing the model's generalization capability, whereas SVM, MLP, and LSTM are more prone to overfitting or are sensitive to noise [14].

For the image and text fusion, this study uses the Text-Image CNN method. Compared with traditional methods such as SVM, MLP, and LSTM, Text-Image CNN can establish a direct spatial-semantic association between image and text features through the convolutional layers, making it especially suitable for tasks that require alignment between textual descriptions and image content [15,16]. In contrast, methods like SVM and MLP lack the ability to capture spatial-semantic associations, while LSTM, though capable of processing sequential text information, has limited capacity to learn spatial information from images [17].

2.3.1. Speech-image feature fusion (MCB)

In a multimodal data fusion framework, fusion of speech and image features is an important step in recognizing nocturnal groaning disorder. In this paper, Multimodal Compact Bilinear Pooling (MCB) method is used for feature level fusion to improve the recognition accuracy.

MCB is an efficient multimodal feature fusion method that captures the complex relationships between different modal features through bilinear pooling while reducing the number of parameters and computational complexity [18]. The fusion process of MCB is divided into three steps:

1) Input features: Speech features are extracted by MFCC and represented as vectors $f_a \in \mathbb{R}^{d_a}$ Input features: Speech features are extracted by MFCC and represented as vectors. Image features are extracted by MobileNetV2 and are represented as vectors $f_i \in \mathbb{R}^{d_i}$ The image features are extracted by MobileNetV2 and represented as vectors.
2) MCB fusion steps:

Step 1: Combine the speech features f_a and image features f_i respectively by two independent random projection matrices $P_a \in \mathbb{R}^{d_a \times d_h}$ and $P_i \in \mathbb{R}^{d_i \times d_h}$ projected into the high-dimensional space to get the high-dimensional features h_a and h_i:

$$h_a = P_a f_a \tag{4}$$

$$h_i = P_i f_i \tag{5}$$

Step 2: Combine the projected high-dimensional feature vectors h_a and h_i multiply element-by-element (Hadamard product) to get the fused feature$h$ and

$$h = h_a \cdot h_i \tag{6}$$

where $\cdot$ denotes the Hadamard product operation, i.e., element-by-element multiplication.

Step 3: Use the Count Sketch technique to map the fused features back to the lower dimensional space h. Mapping the fused features back to the low dimensional space to get a tight fusion feature f_{fusion} Step 4: Mapping

$$f_{fusion} = CountSketch(h) \tag{7}$$

3) Fusion feature representation: the final fusion feature vector obtained $f_{fusion} \in \mathbb{R}^{d_f}$ contains the combined information of speech and image features for subsequent classification tasks.

2.3.2. Image-text fusion (Text-Image convolutional neural network, Text-Image CNN)

In a multimodal data fusion framework, the fusion of image and text features is a key aspect of nocturnal groaning syndrome recognition. In this paper, a Text-Image Convolutional Neural Network (Text-Image CNN) is used for feature-level fusion, combining text features extracted by TF-IDF and image features extracted by MobileNetV2 to obtain a more comprehensive representation of information.

Text-Image CNN utilizes Convolutional Neural Networks (CNNs) to process both text and image features simultaneously, capturing the correlation between the two through a shared convolutional and feature convergence layer to generate a joint feature representation [19]. Its fusion process includes the following steps:

1) Input features: Text features are extracted using TF-IDF, represented as vectors

$$f_t \in \mathbb{R}^{d_t} \tag{8}$$

where d_t is the dimension of the text features. The image features are extracted by MobileNetV2 and are represented as vectors

$$f_i \in \mathbb{R}^{d_i} \tag{9}$$

where d_i is the dimension of the image feature.

2) Text-Image CNN architecture: Shared convolutional and pooling layers process text and image features to capture spatial and semantic relationships.

3) Fusion feature generation: Text and image features are used to generate feature mapping through a shared convolutional layer.

$$c_t = CNN_t(f_t) \tag{10}$$

$$c_i = CNN_i(f_i) \tag{11}$$

Of these, the c_t and c_i are the feature mappings obtained after CNN processing of text and image features, respectively.

2) Feature fusion: combining feature mappings into joint feature vectors through a feature convergence layer.

$$h = Pool([c_t, c_i]) \tag{12}$$

Here the $[c_t, c_i]$ denotes that the feature mappings of text and image are connected by columns, and$Pool$ is the average pooling.

5) Fusion feature representation: The final obtained fusion feature vector:

$$f_{fusion} = h \tag{13}$$

2.3.3. Integration strategies at the decision-making level

Using different neural networks for different channels can maximize the recognition rate for a single channel, while fusion at the decision layer can improve the accuracy of the recognition results. In this paper, the feature fusion part is optimized, in the speech and image channels, MCB fusion technique is used to fuse the speech detail features obtained using MFCC and the image detail features obtained by MobileNetV2, in the image and text channels, the Text-Image CNN network is used to fuse the obtained image detail features and text features at the feature layer and at the decision layer. The recognition results of different channels are fused according to the weighting criterion, and the recognition results are output with the probability on each classification. The weighting criterion is shown in the following equation:

$$p_{fusion} = w_1 p_1 + w_2 p_2 \tag{14}$$

where p_{fusion} is the weighted probability vector for decision layer fusion, and p_1 denotes the probability on the image, speech channel, and p_2 denotes the probability on the image, text channels, the w_1 and w_2 are the weights on the two channels respectively, this paper takes w_1 = 0.7, and w_2 = 0.3.

3. Analysis of experiments and experimental results

3.1. Data set processing

The dataset used for the experiments was provided by Peking University Stomatological Hospital and included three modalities: audio, image and text. The audio dataset contained 1507 recordings, 875 for patients with nocturnal groaning disorder and 632 for non-patients. The image data consisted of 113 face images and 29 brain CT images for nocturnal groaning disorder patients; 12 and 4 images each for patients who also had obstructive sleep apnea syndrome; 144 face images and 48 brain CT images for non-patients; and 60 and 16 images for patients with obstructive sleep apnea syndrome only. Text data are basic patient information. Patients with nocturnal groaning were categorized as experimental group and non-patients as control group for subsequent analysis. The composition of the dataset is shown in **Table 1**.

Table 1. Composition of the data set.

Modal	Classification	Training	Validation	Test	total
Audio	Groaning	612	88	175	875
	Control	442	64	126	632
Image	Groaning	111	16	31	158
	Control	188	27	53	268
Text	Groaning	23	3	7	33
	Control	45	6	13	64
total		1421	204	405	2030

3.2. Experimental details

The experiments in this paper were conducted on Python 3.6 with a hardware platform of Intel® Xeon® Silver 4210 CPU (2.2 GHz), 32 GB of RAM and NVIDIA Quadro P4000 GPU (8 GB). To validate the efficiency of the multimodal fusion technique for nocturnal groaning syndrome recognition, a feature fusion model was constructed including audio (MFCC), image (MobileNetV2) and text (TF-IDF) feature extraction. Multimodal Compact Bilinear Pooling (MCB) was used to fuse speech and image features, Text-Image Convolutional Neural Network (Text-Image CNN) was used to fuse text and image features, and Support Vector Machines (SVMs) were used for pattern recognition. Finally, the recognition results of each modality are merged using weighted fusion method.

The dataset is divided into training, validation and test sets in 70:20:10. Adam optimizer (initial learning rate 0.001), cross-entropy loss function, batch size 32, 50 rounds of training and early stopping strategy to prevent overfitting are used, supplemented with data enhancement techniques. The training process includes model initialization, feature extraction, feature fusion, model training and evaluation.

3.3. Analysis of experimental results

In analyzing audio data for nighttime groaning syndrome, SVM [20] outperforms DBN [21] due to its strengths with small sample sizes. SVM's ability to identify optimal classification boundaries allows it to utilize each data sample effectively, achieving high accuracy even in limited datasets. In contrast, DBN often requires larger datasets to generalize well, as smaller samples may lead to local minima, impacting performance.

For image data, SVM is advantageous over MLP when handling high-dimensional feature spaces. SVM excels in constructing complex classification boundaries by maximizing inter-class distances, which enhances accuracy for image-based tasks. MLP, however, relies heavily on large datasets for effective feature extraction, and limited data can lead it to suboptimal classification [22].

For text data, SVM offers computational efficiency and robustness compared to LSTM [23], which demands longer training times and significant hardware resources due to its iterative weight updates. SVM's simpler model structure also enhances interpretability, making it practical in clinical settings where result transparency is key.

Table 2 counts the comparison of the recognition effect of the unimodal algorithm selected in this paper compared to other algorithms for speech mode, image mode, and text mode.

Table 2. Unimodal recognition effect.

Modal	Feature	Method	Accuracy	AUC
Audio	MFCC	DBN	82.1%	0.88
	MFCC	SVM	85.1%	0.90
Image	MobileNetV2	MLP	80.3%	0.86
	MobileNetV2	SVM	82.5%	0.87
Text	TF-IDF	LSTM	81.2%	0.87
	TF-IDF	SVM	79.3%	0.85

As can be seen from **Table 2**, the recognition accuracy of MFCC features combined with SVM is better than that of DBN in speech modality.SVM is more suitable for dealing with high-dimensional data and small sample sets, and it can effectively utilize the frequency information of MFCC to find the optimal classification hyperplane and improve the accuracy. On the other hand, DBN's performance on small sample sets is unstable, complicated to train, and easy to fall into local optimization. SVM is also superior to DBN in terms of AUC value.

In image modality, the recognition accuracy of MobileNetV2 features combined with SVM is higher than that of MLP.SVM efficiently handles high-dimensional complex features and avoids the overfitting problem of MLP. Although MLP can handle nonlinear problems, it is sensitive to high-dimensional features, complex to train, and weak in generalization.The AUC value shows that SVM is slightly better than MLP [24].

In text modality, the recognition accuracy of TF-IDF combined with SVM is slightly lower than LSTM, but still performs better. The experiment used data from Peking University Stomatological Hospital to ensure the unity of speech and image modalities.

From **Table 3,** it can be seen that the recognition accuracy of this paper's method in speech and image modalities is slightly lower than that of LSTM using ZCR + Hough transform method, but still performs well. the MFCC features efficiently capture the speech spectral information, while MobileNetV2 extracts the deep image features. the SVM performs well in dealing with the high-dimensional data and is able to find the optimal classification boundaries, which improves the accuracy.

In image and text modalities, this paper's method outperforms the other three methods. the effective combination of MobileNetV2 and TF-IDF features improves the recognition performance. **Table 4** shows that speech and image features fused by MFCC and MobileNetV2 can improve the accuracy, while image and text features fused by MobileNetV2 and TF-IDF can also improve the recognition effect, indicating that the multimodal fusion strategy significantly enhances the recognition performance.

Table 3. Comparison of recognition effect on dual modality.

Modal	Feature	Method	Accuracy	AUC
Audio, Image	LPC + SIFT	SVM	87.5%	0.91
	STFT + LBP	MLP	88.4%	0.92
	ZCR + Hough transform	LSTM	91.4%	0.94
	MFCC + MobileNetV2	MCB	89.2%	0.92
Image, Text	SIFT + BoW	SVM	84.6%	0.89
	LBP + LSA	MLP	86.9%	0.90
	Hough transform + word embedding	LSTM	89.3%	0.92
	MobileNetV2 + TF-IDF	Text-Image CNN	89.4%	0.93

Table 4. Comparison of unimodal, bimodal and multimodal recognition effects.

Modal	Feature	Method	Accuracy	AUC
Audio	MFCC	DBN	82.1%	0.88
Image	MobileNetV2	MLP	80.3%	0.86
Text	TF-IDF	LSTM	81.2%	0.87
Audio + Image	MFCC + MobileNetV2	MCB	89.2%	0.92
Image + Text	MobileNetV2 + TF-IDF	Text-Image CNN	89.4%	0.93
Audio + Image + Text	MFCC + MobileNetV2 + TF-IDF	SVM	89.5%	0.94

Table 5. Statistics of test set identification results.

category	Number of samples	correctly identified samples	Accuracy	AUC
Groaning	213	190	89.2%	0.93
Control	192	171	89%	0.93
(grand) total	405	361	89.5%	0.94

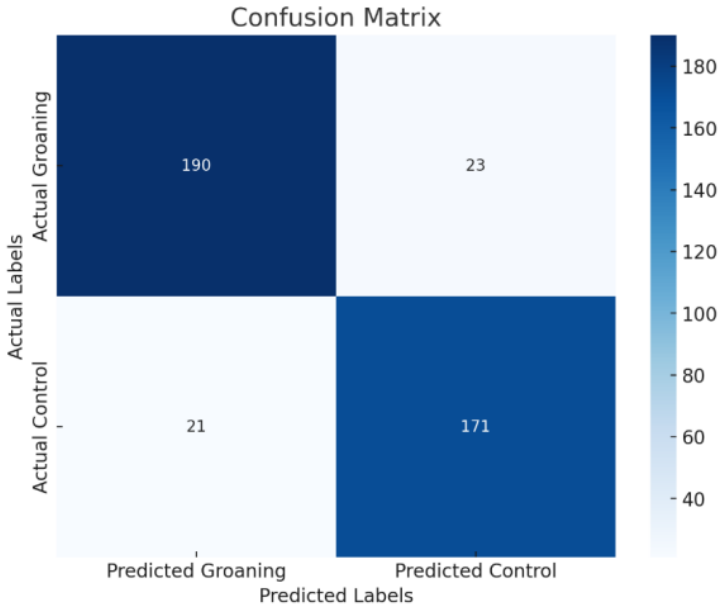

Figure 4. Confusion matrix of multimodal identification results.

The recognition accuracies of the experiments on the samples of the three modalities are shown in **Table 5**, and the confusion matrices of the recognition results on the two categories on the test set are shown in **Figure 4**. The horizontal coordinates of the confusion matrix represent the results of the predicted types and the vertical coordinates represent the type distribution of the real samples. When the horizontal and vertical coordinates are consistent, it represents correct identification, and when they are not consistent, it means that the type of the real sample represented by the vertical coordinate has been incorrectly predicted to be another type; the confusion matrix is more visual, so that we can see the distribution of the samples on each type, and each confusion matrix represents a type of identification result, which is a supplement to this identification statistics table. As can be seen from **Table 5** and **Figure 4**, both types achieved good results, with an overall recognition accuracy of

89.5%, which is an improvement in recognition accuracy compared to traditional unimodal.

4. Conclusion

This paper proposes a multimodal nighttime groaning recognition method based on speech, image, and text, and conducts related experiments. The results show that the multimodal fusion strategy significantly improves the recognition accuracy of nighttime groaning compared to traditional single-modal methods. Although the model performs well in an experimental setting, real-world application scenarios often involve more complex environmental noise and inconsistent data, which may impact model performance. For instance, when patients use this system in a home environment, the quality of audio and video capture devices may vary significantly. Therefore, further improvements to the network are needed to enhance the differentiation of similar features.

Author contributions: Conceptualization, XX and WS; methodology, XX; software, XX; validation, XX, WS and YL; formal analysis, XX and YL; investigation, XX; resources, XG (Xu Gong); data curation, XG (Xuemei Gao) and MY; writing—original draft preparation, XX; writing—review and editing, QW and YL; visualization, XG (Xu Gong); supervision, QW and XG (Xuemei Gao); project administration, QW and XG (Xuemei Gao); funding acquisition, QW. All authors have read and agreed to the published version of the manuscript.

Ethical approval: The study protocol complies with the principles and requirements of the Declaration of Helsinki and has been approved by the Biomedical Ethics Committee of Peking University Hospital of Stomatology (approval number: PKUSSIRB-201631128).

Informed consent: Informed consent was obtained from all subjects involved in the study.

Funding: National Natural Science Foundation of China (grant No. 81670082).

Conflict of interest: The authors declare no conflict of interest.

References

1. Oudiette, D, LeuSemenescu, et. al. Nocturnal groaning: an unusual sleep-related vocalization.[J]. Sleep Medicine, 2018(41):7-13.
2. Ferri, R, Manconi, et. al. Nocturnal groaning: sleep-related disorders, singing, and nocturnal vocalizations[J]. Handbook of Clinical Neurology, 2017(146):289-297.
3. Mallat, S. A Theory for Multiresolution Signal Decomposition: The Wavelet Representation. [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1989(11(7)):674-693.
4. Donoho, L D. Denoising by Soft-Thresholding[J]. IEEE Transactions on Information Theory, 1995(41(3)):613-627.
5. Zhang, X, & Desai, et. al. Adaptive Denoising Based on SURE Risk[J]. IEEE Signal Processing Letters, 2003(10(4)):113-116.
6. Gonzalez, C R, & Woods, et. al. Digital Image Processing[M]. 2nd ed. Prentice Hall, 2002.
7. Bovik, A. C. (Ed.). Handbook of Image and Video Processing[M]. Academic Press, 2005.
8. Sun Xiaohu, Li Hongjun. A review of speech emotion recognition[J]. Computer Engineering and Applications, 2020(56(11)):1-9.

9. HOU Wenqi, YANG Shihua, WU Zhifeng, et al. Search and optimization of MobileNetV3[J]. Computer Science and Exploration, 2019(13(4)):567-580.
10. Sharmandin, Hou WQ, Zhu M, et al. MobileNetV2: Inverse residuals and linear bottlenecks[C]: Proceedings of the Conference on Computer Vision and Pattern Recognition, 2018.
11. Ghosh, A, Bhattacharya, etal. AdGAP: Advanced Global Average Pooling[C]. //Proceedings of the 2018 International Conference on Machine Learning and Data Engineering (ICMLDE), 2018:16-21.
12. Shalton, Michael. Introduction to Modern Information Retrieval [M]. Beijing: Tsinghua University Press, 1986.
13. Fukui, A, Park, et al. Multimodal compact bilinear pooling for visual question answering and visual grounding[J]. Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, 2016:457-468.
14. Karpathy, A, & Fei-Fei, et al. Deep visual-semantic alignments for generating image descriptions[J]. EEE Transactions on Pattern Analysis and Machine Intelligence, 2015(39(4)):664-676.
15. LeCun, Y, Bottou, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998(86(11)):2278-2324.
16. Simonyan, K, & Zisserman, et al. Very deep convolutional networks for large-scale image recognition[C]. //Proceedings of the International Conference on Learning Representations, 2015.
17. Sainath, N T, Mohamed, etal. Deep Convolutional Neural Networks for LVCSR[J]. Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, 2013:8614-8618.
18. Gao HB, Zhang N, Deng ZC, et al. Compact bilinear pooling[C]: Proceedings of the Conference on Computer Vision and Pattern Recognition, 2016.
19. XU Kaiyuan, BAI Zhilong, WU Donghui, et al. Visual attention mechanisms in image subtitle generation[C]: Proceedings of the International Conference on Machine Learning, 2015.
20. Cortes, C, & Vapnik, et al. Support-vector networks[J]. Machine Learning, 1995(20(3)):273-297.
21. Hinton, E G, Osindero, et al. A fast learning algorithm for deep belief nets. [J]. Neural Computation, 2006(18(7)):1527-1554.
22. Bishop, M C. Pattern Recognition and Machine Learning[M]. Springer, 2006.
23. Hochreiter, S, & Schmidhuber, et al. Long short-term memory[J]. Neural Computation, 1997(9(8)):1735-1780.
24. Vapnik, V. Statistical Learning Theory[M]. Wiley, 1998.