

International Journal of Simulation and Process Modelling

ISSN online: 1740-2131 - ISSN print: 1740-2123

<https://www.inderscience.com/ijspm>

**Multimodal simulation for diagnosing microstrabismus
integrating electrooculogram and electroencephalographic
signals**

Tong Yang, Shaoyun Bai

DOI: [10.1504/IJSPM.2026.10079108](https://doi.org/10.1504/IJSPM.2026.10079108)

Article History:

Received:	08 December 2025
Accepted:	17 May 2026
Published online:	01 October 2026

Multimodal simulation for diagnosing microstrabismus integrating electrooculogram and electroencephalographic signals

Tong Yang and Shaoyun Bai*

College of Computer Science and Technology,
Shanxi University of Electronic Science and Technology,
Linfen, 041000, China
Email: yangtong@sxdzkj.edu.cn
Email: baishaoyun@sxdzkj.edu.cn
*Corresponding author

Abstract: Microstrabismus is difficult to diagnose using conventional methods due to its subtle presentation and high risk of misdiagnosis. To overcome this limitation, we propose a multimodal diagnostic framework that integrates electrooculogram (EOG) and electroencephalogram (EEG) signals. The framework first preprocesses both signal types and applies principal component analysis to remove redundancy while preserving essential information. A dual-branch architecture then extracts spatio-temporal features from each modality, which are subsequently fused through a multi-layer interaction mechanism designed to capture cross-modal complementarity. To handle inter-sample variability, a gating-based module dynamically adjusts the fusion ratio between modalities according to individual sample characteristics. Experimental results indicate that the proposed model improves diagnostic accuracy by at least 6.35% over baseline methods, demonstrating strong potential for aiding the precise diagnosis of micro-degree strabismus.

Keywords: strabismus diagnosis; multimodal feature fusion; EOG signal; EEG signal; attention mechanism.

Reference to this paper should be made as follows: Yang, T. and Bai, S. (2026) 'Multimodal simulation for diagnosing microstrabismus integrating electrooculogram and electroencephalographic signals', *Int. J. Simulation and Process Modelling*, Vol. 23, No. 3, pp.169–183.

Biographical notes: Tong Yang is a Lecturer in the College of Computer Science and Technology at Shanxi University of Electronic Science and Technology, China. He received his Bachelor's degree from Changchun Normal University (2009) and Master's degree from Yunnan University (2012), China. He published two papers. His research interests are software engineering, machine vision and brain-computer interface.

Shaoyun Bai is a Lecturer in the College of Computer Science and Technology at Shanxi University of Electronic Science and Technology, China. He received his Bachelor's degree in XinZhou Normal University (2013) and Master's degree in Taiyuan University of Science and Technology (2016), China. He published three papers. His research interests are database, software engineering and big data analysis.

1 Introduction

Strabismus is a common visual impairment characterised by misalignment of the visual axes, resulting in abnormal binocular vision function (Bui and Milleret, 2014). Among various types of strabismus, microstrabismus poses a unique diagnostic challenge due to its subtle clinical presentation and susceptibility to being overlooked in routine examinations. Traditional diagnostic methods lack sufficient sensitivity for detecting such minute ocular deviations (Yehezkel et al., 2020). Subjective perception tests, which rely on patient cooperation and feedback, face limitations in application among special populations such as

children (Araújo Santos et al., 2025). With the advancement of bioelectric signal acquisition technology, electrooculogram (EOG) and electroencephalogram (EEG) signals have opened a new window for elucidating the neurophysiological mechanisms underlying microstrabismus. EOG signals can precisely record changes in the corneal-retinal potential difference, thereby capturing minute ocular movement trajectories with high accuracy, making it an ideal tool for objectively quantifying ocular position (Yang et al., 2025). Physiologically, EOG reflects the peripheral oculomotor output, directly measuring the misalignment angle. However, EOG alone can only reflect the output of eye movements and cannot reveal the

perceptual process of the visual cortex in responding to binocular visual conflicts (Latifoglu et al., 2020). EEG signals, on the other hand, directly reflect the electrophysiological activity of the cerebral cortex, particularly in the occipital and parietal lobes associated with visual processing. When microstrabismus occurs, the visual information transmitted from both eyes exhibits a slight spatial misalignment. To maintain fusion, the brain must exert greater neural effort, potentially generating specific cortical activity modulation (Chen et al., 2023). Therefore, integrating EOG with EEG to establish a multimodal assessment framework holds promise. This combination ensures that both the peripheral motor deviation EOG and the central sensory adaptation EEG are captured, providing a comprehensive biomarker for microstrabismus.

The early diagnostic methods for mild strabismus were mainly based on the extracted physiological signal features through manual processing, and were classified and discriminated by statistical methods or simple machine learning models. However, due to the subjectivity of manual operations and the limitations of the models, there were significant shortcomings in the diagnosis of extremely mild strabismus, the consistency of diagnosis, and efficiency. These diagnostic methods for microstrabismus relied on manually extracting time-domain, frequency-domain, and nonlinear features from EOG signals, combined with traditional classification algorithms such as support vector machine (SVM) (Das et al., 2024), random forests (Antoniou et al., 2021), and K-nearest neighbours (Ghosh et al., 2023) for strabismus diagnosis. Cavaliere et al. (2019) employed empirical mode decomposition to decompose EOG signals, utilising sample entropy as the feature for analysis. They applied SVM for vision diagnosis, achieving an accuracy rate of 78.98%. Matsui et al. (2022) employed a fractal pattern feature extraction method using TQWT to extract electrooculographic features, achieving a classification accuracy of 79.82% using the SVM algorithm. Tanya et al. (2023) extracted statistical and frequency-domain features related to pupil diameter and fixation distance from EOG signals, applying decision trees for visual acuity diagnosis with high diagnostic accuracy. Meidelfi et al. (2023) decomposed EEG signals using discrete wavelet transform (DWT) to extract features across multiple frequency bands, then employed the k-nearest neighbours algorithm for classification, ultimately achieving an accuracy rate of 83.26%. Abbas et al. (2022) analysed transitions from relaxed to stressed states in computer users to extract pupil diameter features for vision diagnosis, achieving an average accuracy rate of 82.21%.

In the early days of machine learning, features were extracted entirely based on manual experience. Diagnosticians needed to combine the pathological mechanisms of strabismus. Features were manually defined and extracted from the raw EEG/EOG signals, such as the amplitude difference of EEG waves, latency, the eye position offset of EOG, and the movement synchronicity, etc. The effectiveness of the features entirely depended on

the professional knowledge of the diagnosticians, and only shallow, interpretable features could be extracted, unable to capture the hidden complex correlations in the signals. Deep learning methods achieve end-to-end automated feature extraction. Without the need for manual intervention, neural networks can directly take the raw EEG/EOG signals as input and, through convolution, recurrent or attention mechanisms in multiple layers of the network, gradually explore the deep features in the signals. By constructing multi-layer neural networks, these methods can automatically extract high-level features from raw electrooculographic signals, reducing reliance on manual feature design. Ahmed et al. (2021) proposed a strabismus diagnosis model based on the multi-column convolutional neural network (CNN), which integrates the outputs from multiple recognition modules and employs a weighted sum strategy to generate the final decision. Muthukannan (2022) proposed a recurrent convolutional network model combining CNN and recurrent neural network (RNN), where CNN primarily extracts spatial features from EOG data while RNN captures temporal information. This study classified five types of strabismus (esotropia, exotropia, vertical strabismus, rotational strabismus, and mixed type) on a self-built dataset. The experimental results showed that the model achieved an average classification accuracy of 86.2%. However, the dataset used in this study only included 40 subjects, and no comparison verification with public datasets was conducted. Therefore, its generalisation ability still needs to be further tested. Varošanec et al. (2024) proposed a strabismus diagnosis model based on bidirectional long short-term memory network (Bi-LSTM), which combines forward LSTM and backward LSTM to form the final Bi-LSTM architecture, ultimately tested on the AMIGOS dataset. El-Ateif and Idri (2024) introduced a deep belief network, extracting differential entropy features as input. Subsequently, integrating a hidden Markov model achieved an average accuracy of 84.62% on the SEED dataset. This dataset contains 15 subjects' EEG recordings when they were watching emotion-evoking videos, but does not include EOG signals. Therefore, the clinical applicability of this method still needs to be further verified in scenarios that include eye movement data. Qian et al. (2025) employed graph convolutional network (GCN) and LSTM for strabismus diagnosis. This study achieved a diagnostic accuracy rate of 89.3% on a self-built EOG dataset consisting of 80 participants. Among them, the recognition effect for horizontal strabismus was the best. However, the recognition rate for rotational strabismus was only 84.7%, indicating that a single EOG modality still has a bottleneck in recognising complex strabismus types.

Single-modality strabismus diagnostic methods exhibit significant variations in their ability to distinguish between different types of strabismus. Consequently, many researchers have gradually shifted their focus toward multimodal strabismus diagnostic approaches. Multimodal fusion integrates information from multiple data sources, enabling more reliable prediction of patterns for the same phenomenon and effectively capturing complementary

information across different modalities. Zhang et al. (2023) developed a multimodal feature fusion neural network (MFFNN) to effectively extract complementary information from EOG signals and achieve deep integration with EEG signals, thereby enhancing diagnostic accuracy. Wu et al. (2025) constructed a multimodal strabismus diagnosis framework based on swarm intelligence and CNN. This study employed EEG as the core physiological indicator, effectively integrating it with EOG data. During feature extraction, the researchers utilised multivariate synchronous compressive transform technology to obtain discriminative features from different modalities. Garg et al. (2025) employed a CNN to learn the spatial representation of multi-channel EEG signals and an LSTM to learn the temporal representation of EOG signals, combining both representations for strabismus diagnosis tasks. Joshi (2025) proposed a multimodal strabismus diagnosis model integrating multi-scale feature representations and attention mechanisms. This model extracts EEG features through a multi-stream network and employs a dual-scale attention module for effective EOG feature extraction. The model achieved a diagnostic accuracy of 86.2%.

In summary, current research has not fully leveraged the complementary information from multimodal physiological signals, resulting in suboptimal diagnostic accuracy. To address these issues, this paper proposes a diagnostic model for microstrabismus that integrates EOG and EEG signals, termed the multimodal diagnostic model for microstrabismus (MDMS). The contribution of this study lies in constructing an interpretable multimodal diagnostic framework, where each component corresponds to different stages of the clinical diagnostic process. Firstly, the dual-branch feature extraction module simulates the ophthalmologist's independent assessment of EOG and EEG. Secondly, the cross-modal interactive attention mechanism simulates the clinical reasoning process of integrating peripheral eye movement behaviour with central neural processing information. Thirdly, the gating weighted fusion module simulates the decision-making process where clinicians dynamically adjust the diagnostic weights of each modality based on individual patient differences. Through this architectural design, the MDMS model not only achieves high-precision classification but also provides an interpretable diagnostic decision framework that helps understand the complementary role of multimodal physiological signals in microstrabismus diagnosis.

The main work of this model can be summarised in the following aspects.

- 1 Perform preprocessing and normalisation on EOG and EEG signals. Employ principal component analysis (PCA) to reduce the dimensionality of the original EOG and EEG signal data, thereby eliminating redundant information while preserving the primary information of the original signals.
- 2 The spatio-temporal features of EOG and EEG signals are extracted based on a dual-branch architecture, and efficiently fused through a multi-layer interaction

mechanism. Unlike most existing approaches that employ static concatenation or simple fusion, the cross-modal interactive attention mechanism proposed in this paper establishes deep dependencies between the EEG and EOG modalities. This enables the model to capture complementary information across modalities with greater flexibility and effectiveness.

- 3 Considering that the importance of different modalities may dynamically vary across samples, this paper designs a gating mechanism-based modal weight adjustment module. This module automatically adjusts the fusion ratio between EEG and EOG modalities during training based on sample features, thereby enhancing adaptability to marginal samples while maintaining overall stability.
- 4 Multimodal simulations were conducted on the proposed model. Experimental results demonstrated that the proposed model achieved diagnostic accuracy and F1 scores of 93.41% and 93.72%, respectively, outperforming the comparison model. The proposed model holds significant practical value in enhancing diagnostic efficiency and reducing misdiagnosis rates, thereby supporting early intervention in visual health.

2 Relevant theoretical analysis

2.1 Attention mechanism

An attention mechanism is a technique used to enhance the focus of neural networks on important information, thereby improving model performance (Zhang and Cheng, 2024). The attention mechanism consists of a query matrix (*query*), key (*key*), and weighted average, forming a multi-layer perceptron (MLP) (Ma et al., 2019). Given an element *query* in target, by calculating the similarity or relevance between *query* and each *key*, the weight coefficient corresponding to each *key*'s value is obtained. Then, the values are weighted and summed to obtain the final attention value. According to its principle, the calculation process is as follows, where L_x is the length of source, and $L_x = \|Source\|$.

$$\begin{aligned} & \text{Attention}(Query, Source) \\ &= \sum_{i=1}^{L_x} \text{Similarity}(Query, Key_i) \cdot Value_i \end{aligned} \quad (1)$$

The attention mechanism calculation process includes three stages: similarity computation, weight normalisation, and context vector generation. Similarity computation refers to calculating the matching degree between *query* and *key*. Weight normalisation involves normalising the original weights obtained in the similarity computation stage, mapping the numerical range to between 0 and 1, as shown in equation (2), where $\alpha_{i,j}$ is the attention weight of the i^{th} query for the j^{th} key.

$$\alpha_{i,j} = \text{Softmax}(\text{score}(Q_i, K_j)) = \frac{\exp(\text{score}(Q_i, K_j))}{\sum_k \exp(\text{score}(Q_i, K_k))} \quad (2)$$

The attention mechanism enables the model to automatically assign weights to different features based on the characteristics of the input data. For instance, in the diagnosis of strabismus, the EOG signal may contain noise such as eye movements and blinking, while the EEG signal may reflect the correlation between brain activity and strabismus. The attention mechanism can focus on the key features related to strabismus and suppress irrelevant noise, thereby extracting purer and more discriminative features.

2.2 Multimodal fusion strategy

The diagnostic method based on a single modality for strabismus provides relatively limited information, which may not fully reflect the patient's strabismus condition. For instance, using only the EOG signal can only capture information about eye movement, but it cannot provide insights into the patient's brain activity or the structure of the eyes. The multi-modal strabismus diagnostic method, by integrating multiple data sources, can offer more comprehensive and detailed information. For example, combining the EOG signal with the EEG signal can simultaneously understand eye movement and brain activity, and combining eye images can observe abnormal eye structures, thereby providing a more accurate diagnosis of strabismus. Multimodal fusion strategies aim to integrate information from different modalities to fully leverage their respective strengths and compensate for the limitations of any single modality, thereby enhancing overall system performance. Multimodal data exhibit distinct complementarity and interconnectivity; effective fusion strategies can significantly improve the accuracy and robustness of downstream tasks such as recognition, prediction, or generation (Li et al., 2025). This section will explore the core concepts and common approaches to multimodal fusion, focusing on three dimensions: feature-level fusion, decision-level fusion, and hybrid fusion.

- 1 **Feature-level fusion.** Its core concept involves directly merging features from different modalities after data preprocessing to construct a unified joint feature space (Nie et al., 2021). The advantage of this strategy lies in its intuitiveness and efficiency. Since fusion occurs early in the feature extraction process, information loss from raw data is minimised, and the model can directly learn latent correlations between different modalities. Feature-level fusion belongs to early fusion. Any noise or distortion in a single modality data will be directly transmitted to the fused features, thereby affecting the overall diagnostic effect. Moreover, if the feature extractor of a certain modality fails to learn effective pathological features, the fused feature set will have inherent defects, and the subsequent classifier cannot make up for them. At the same time, the structure of the

feature-level fusion model is fixed, making it difficult to adapt to the feature differences of different strabismus types, and the generalisation ability is limited.

- 2 **Decision-level fusion.** This strategy emphasises independent feature extraction and discriminative inference for each modality, ultimately synthesising the outputs or probabilities from all modalities at the decision layer to produce a comprehensive fusion decision (Pawłowski et al., 2023). Its advantages lie in flexibility and robustness. Since classifiers for each modality are trained independently, decision-layer fusion effectively avoids feature interference between modalities. Decision-level fusion is prone to being interfered by the erroneous decisions of a certain unreliable modality. Even if a meta classifier is introduced, a significant amount of intermediate evidence that could be used to judge the reliability is lost because the input is only the high-level abstract decision.
- 3 **Hybrid Fusion.** This strategy introduces cross-modal interaction mechanisms during the mid-stage of feature extraction, balancing the information completeness of early fusion with the flexibility of late fusion (Zhao et al., 2024). Mixing and fusion require an enormous amount of labelled data for effective training; otherwise, overfitting is highly likely to occur. In fields with extremely high labelling costs such as healthcare, this often becomes the main bottleneck for their application. At the same time, the computational and storage costs also significantly increase.

3 Preprocessing of EOG and EEG signals

3.1 Preprocessing process

In the research of multimodal strabismus diagnosis models, data preprocessing is a critical step, as its quality directly impacts the effectiveness of downstream feature extraction and model training. When dealing with both EEG and EOG signals, further cleaning, synchronisation, slicing, and normalisation of raw or semi-processed data are also required.

In the EOG signal processing stage, differentiated imputation strategies are applied to address missing values in the dataset based on their underlying causes. For conventional missing values, the mean of the preceding and subsequent data points is used for imputation, effectively preserving the overall trend of the data. For missing pupil diameter data caused by blinking, linear interpolation algorithms are employed for imputation. This method achieves more precise reconstruction of data characteristics during blinking by analysing trends in data changes before and after the blink. To ensure temporal alignment of multimodal data, the EOG signal undergoes upsampling. This ensures the timing of the EOG signal remains

consistent with the EEG signal, laying the foundation for subsequent multimodal feature fusion.

Due to the inherent complexity of physiological signals, collected EEG signals rarely exist in a pure form. They are typically contaminated with noise or artifact signals, including power-line interference, baseline drift, and ECG signals. In traditional EEG-based research methodologies, EOG signals are generally regarded as severe interference, particularly when EEG signals are extracted from the forehead region.

3.2 PCA and normalisation

After preprocessing the EEG and EOG signals, this paper uses PCA (Tharwat, 2016) to reduce the dimensionality of the original data while retaining its main information and removing redundant information. The most important role of PCA is to reduce the dimensionality of signal data; certain elements in multidimensional vectors composed of multiple features have very small variations across the features, making them unsuitable for distinguishing features. In this study, PCA is applied to the raw 62-channel EEG and 33-channel EOG data. This paper retained the top principal components that cumulatively accounted for 95% of the total variance. This process reduced the feature dimension from 95 original channels to 24 principal components, effectively removing 74.7% of redundant information while preserving critical diagnostic signals. The impact of PCA is twofold. Firstly, reducing input dimensions decreases the training time by approximately 40%. Secondly, noise suppression: components with low variance, often corresponding to high-frequency noise or baseline drift, are discarded. This paper extracts five indices from EEG and EOG signals using PCA to reduce dimensions and extract main signal variables: the average power ratio of α waves and β waves in the EEG signals $R_{\alpha\beta}$, the low-to-high frequency average power ratio in horizontal EOG signals $R_{l/h}$, blink frequency, blink duration, and blink speed from the blink features of vertical EOG signals.

Before performing PCA, these features are first normalised. The feature data of the EOG and EEG signals are normalised to $(-1, 1)$ for further processing. The normalisation calculation equation y_{norm} is as follows.

$$y_{norm} = ((y - y_{min}) / (y_{max} - y_{min})) \times 2 - 1 \quad (3)$$

If $y_{max} = y_{min}$, the normalisation equation simplifies to equation (4).

$$y = \frac{y}{y_{max}} \quad (4)$$

After normalising the extracted signal variables in this paper, cluster analysis models from PCA are used to process these signal variables further. The steps for calculating the comprehensive weights of principal components and the comprehensive evaluation model are given below.

- 1 The original data signals must first be standardised. This is done to enable comparability among the various signals. After standardisation, samples are denoted by the following formula, where x is an EEG or EOG signal, $\bar{x}_j$ is the mean of the signal variables, and x_{ij}^* is the variance. Calculate a matrix $R = (r_{ij})_{J \times J}$ related to the indicators, in which $r_{ij} = \sum_{t=1}^N x_{ti}^* x_{tj}^*$.

$$X^* = (x_{ij}^*)_{N \times J} \quad (5)$$

$$x_{ij}^* = \frac{(x_{ij} - \bar{x}_j)}{\sqrt{\frac{1}{N-1} \sum_{t=1}^N (x_{tj} - \bar{x}_j)^2}} \quad (6)$$

$$\bar{x}_j = \frac{\sum_{t=1}^N x_{tj}}{N} \quad (7)$$

- 2 Determine J characteristic roots $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_J \geq 0$ from the symmetric matrix R , the corresponding unit eigenvectors are referred to as the J^{th} principal component of sample X_i .
- 3 Analyse the comprehensive weights of principal components: To extract and summarise the most representative features that characterise the original data, it is common practice to base the selection on the variance contribution rate so that the first p principal components account for more than 85% of all information. This transforms the weight analysis of principal components into a comprehensive evaluation of the variance contribution rates.

When using EEG and EOG to study microstrabismus diagnosis models, the main components most closely related to strabismus diagnosis can be extracted to improve diagnostic accuracy.

4 Multimodal microstrabismus diagnosis model integrating EOG and EEG signals

4.1 Model overview

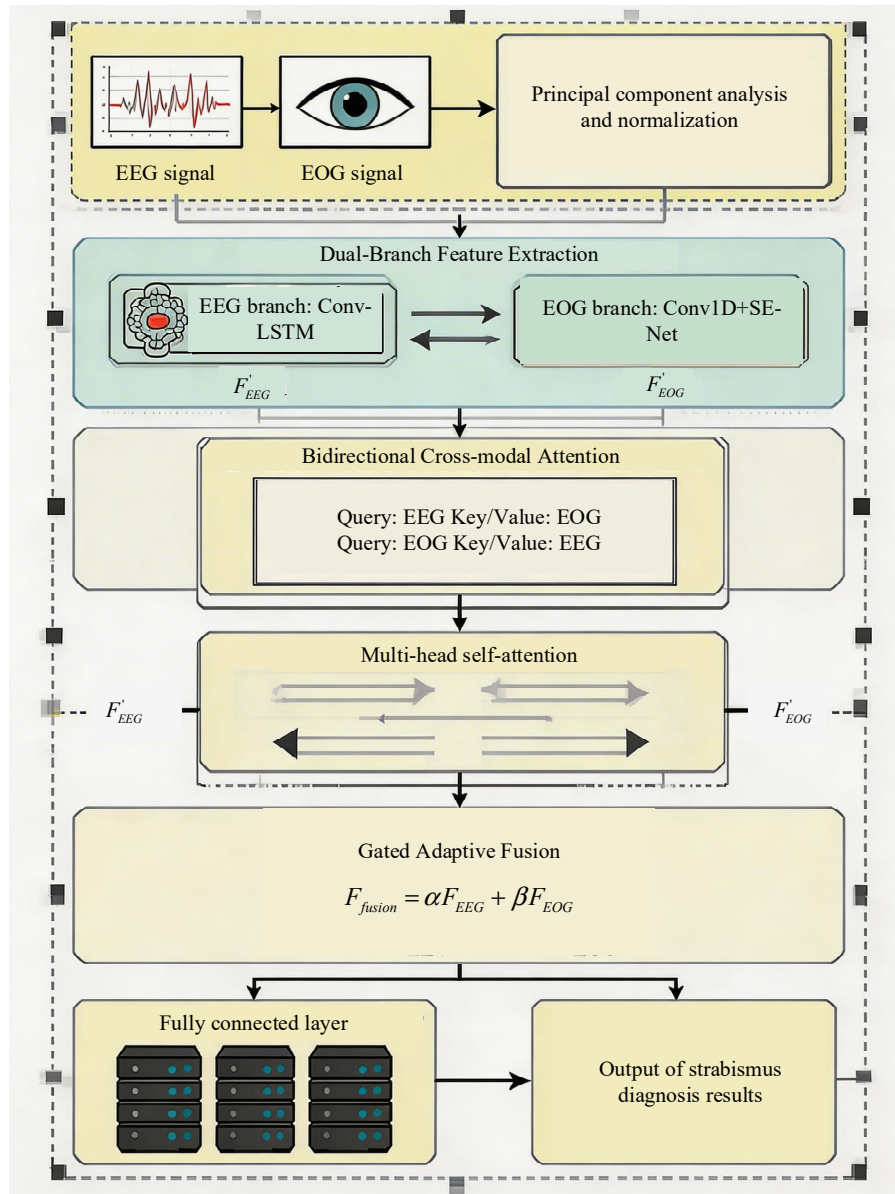
Due to the complexity and diversity in diagnosing microstrabismus, single-modality signals struggle to comprehensively capture its characteristics. Addressing this challenge, recent research has increasingly shifted toward multimodal strabismus diagnosis, integrating multiple physiological signals to enhance diagnostic accuracy and robustness. Although EOG signals are susceptible to environmental and individual variations, they offer the advantage of convenient acquisition. EEG signals, serving as objective indicators of emotional states, provide precise insights despite the complexity of data collection. Traditional methods typically employ independent models to process different modalities, potentially overlooking their

inherent correlations. As shown in Figure 1, this study proposes a multimodal micro-degree strabismus diagnostic framework (MDMS) that integrates EOG and EEG signals. The framework is designed to mirror the clinical diagnostic reasoning process, where multiple physiological indicators are evaluated and integrated to reach a diagnostic conclusion.

1 MDMS performs feature extraction separately for EEG and EOG signals. For EEG signals, MDMS employs a convolutional long short-term memory network to effectively capture spatial structural information and temporal sequence variation features within EEG data. Simultaneously, for EOG signals, the model utilises SENet for feature extraction, highlighting spatial and temporal features within EOG data that significantly contribute to the strabismus diagnosis task.

- 2 After extracting EEG and EOG signal features separately, the algorithm employs a cross-modal interaction attention mechanism to interactively fuse these two modal features, generating a cross-modal feature representation. This cross-modal attention mechanism enables each modal feature to perceive and effectively utilise key information from the other modality, thereby enhancing the recognition capability of the fused features.
- 3 Further leverage the multi-head self-attention mechanism to perform in-depth feature refinement on cross-modal fused features, thereby capturing more complex higher-order interactions between different features. This enhances the discriminative power and robustness of the fused feature representations.

Figure 1 The structure of the suggested micro-degree strabismus diagnosis model MDMS (see online version for colours)



- 4 The MDMS constructs a state vector by feeding multimodal fusion features into a gating mechanism-based adaptive weighting module, dynamically learning the respective fusion weights for EEG and EOG signals. This approach adaptively assigns reasonable weight coefficients to each modality's features through end-to-end backpropagation, thereby generating adaptive fusion features.
- 5 Weighted fusion of EEG and EOG cross-modal attention features is performed using modal weight coefficients obtained through dynamic learning, yielding more precise adaptive fusion features. The fused features are fed into a fully connected classifier for strabismus diagnosis prediction, ultimately calculating the overall diagnostic accuracy rate for micro-degree strabismus.

In summary, the architecture design of MDMS clearly adheres to the principle of simulatability for clinical diagnostic reasoning. Each sub-component of the model is designed to simulate the cognitive operations of ophthalmologists in the actual diagnostic and treatment process.

The dual-branch feature extraction module (step 1) simulates physicians independently reading the EOG eye movement trajectory graphs and the EEG brain waveforms, capturing the abnormal patterns related to microstrabismus from these graphs. The symmetrical structure of this module ensures that each of the two modalities receives a thorough and unbiased primary interpretation before being fused in the subsequent step, similar to when the same physician (or different physicians) independently evaluate different types of examination reports in clinical practice.

Cross-modal interaction attention mechanism (step 2) simulates the advanced integrated reasoning process of physicians. When a physician observes EOG indicating a slight eye position deviation, they will actively backtrack the visual cortex activity within the corresponding time window in the EEG to determine if there is any abnormal fusion compensation in the brain. Conversely, when abnormal alpha wave modulation is observed in the EEG, the physician will re-examine whether there are unnoticed micro-saccades in the EOG. This two-way 'query-key' attention relationship is essentially a mathematical simulation of the clinical cross-validation thinking.

Multi-head self-attention feature refinement (Step 3) simulates the thinking process of physicians after obtaining preliminary integrated evidence, where they repeatedly verify and focus on multiple sources of information. Through the parallel computation of multiple attention heads, the model can explore high-order interaction patterns between modalities from different perspectives (such as the time dimension, spatial electrode dimension, and frequency domain dimension), which is consistent with the practice of clinical experts in comprehensively judging from multiple pathophysiological perspectives.

The adaptive weighted fusion module based on the gating mechanism (steps 4 and 5) simulates the

decision-making behaviour of physicians in adjusting diagnostic weights flexibly in different patients and different clinical situations. In clinical practice, for some patients with eyelid spasms or poor cooperation, their EOG signals may contain a large amount of blinking artifacts. At this time, physicians will naturally reduce their reliance on EOG results and instead trust EEG indicators more. Conversely, when EEG is interfered by electromyography, physicians will rely mainly on EOG. The gating network of MDMS learns end-to-end and automatically calculates dynamic weights (α, β) for each sample, which essentially is a parameterisable simulation of the above dynamic decision-making strategy, and this simulation process has individual-level adaptability.

Through the above design, MDMS is no longer an input-output black box classifier, but a clinical reasoning simulation system with a structurally mappable and behaviourally replicable structure. This simulation framework not only achieves high-precision diagnosis, but more importantly, provides physicians with a traceable and explainable intermediate evidence chain, laying a methodological foundation for subsequent clinical verification and the construction of digital twin systems.

4.2 Dual-branch feature extraction module

This paper proposes a dual-branch feature extraction module with symmetric structure. To reduce computational load, both modalities' input samples are downsampled along the time dimension to 240, where EEG is $1 \times 62 \times 240$ (62 electrodes) and EOG is $1 \times 33 \times 240$.

The symmetric dual-branch architecture is designed to preserve the temporal correlation structures inherent in both EOG and EEG signals while allowing modality-specific feature learning. Temporal correlation in this context refers to the time-dependent relationships between consecutive signal samples, which capture dynamic patterns such as EOG and EEG. The symmetric structure ensures that both modalities undergo comparable temporal feature extraction processes, facilitating effective cross-modal fusion in later stages. Specifically, the temporal window is set to $T = 240$ time steps, which captures sufficient temporal context for microstrabismus-related patterns while maintaining computational efficiency.

- 1 EOG feature extraction. This paper first represents original EOG signal features as a time-varying multidimensional feature matrix, denoted as $X_{EOG} \in R^{T \times D_{EOG}}$, where T is the length of the time series and D_{EOG} is the dimension of EOG features, including multiple indicators such as gaze duration, pupil diameter, saccade amplitude, etc. The temporal correlation in EOG signals primarily manifests as the sequential dependency of eye movement patterns, including saccade velocity profiles, fixation stability over time, and blink-induced signal transitions. To capture these local temporal patterns in EOG data, a one-dimensional CNN (Conv1D) is first applied for preliminary feature extraction. The Conv1D layer uses

a kernel size of 5, stride of 1, and 64 output filters, followed by batch normalisation (BN).

$$U = \text{Conv1D}(X_{EOG}(t)), U \in R^{T \times C} \quad (8)$$

where U represents the feature representation after convolution mapping and $C = 64$ indicates the number of output feature channels used to represent abstract features from different dimensions of EOG signals. The Conv1D layer employs a kernel size of 5 with stride 1, enabling the model to capture temporal dependencies within a 50ms window. This window size is selected based on the typical duration of micro-saccades and fixation micro-tremors associated with strabismus. Subsequently, to highlight channel importance, a visual attention mechanism, specifically the squeeze-and-excitation network (SE-Net), is introduced. SE-Net is a classic implementation of the channel attention mechanism. It compresses the spatial features of each channel into a descriptor through global average pooling. Then, two fully connected layers are used to learn the nonlinear dependencies between channels and generate weight coefficients. Finally, the weights are multiplied by the original feature map channel by channel to enhance the useful features and suppress the redundant information. This network significantly improves the model's perception ability of key features with a relatively low computational cost. First, global average pooling is applied through compression operations on information within each channel to obtain global feature descriptions for all feature channels.

$$Z_c = F_{sq}(U_c) = \frac{1}{T} \sum_{t=1}^T U_c(t), Z_c \in R, c = 1, 2, \dots, C \quad (9)$$

where Z_c is the global descriptive vector for the c^{th} channel.

Through the above operations, the feature values of each channel across all time steps are averaged into a global descriptive vector $Z = [Z_1, Z_2, \dots, Z_C]$.

Subsequently, two fully connected layers perform nonlinear transformations aimed at learning and capturing interactions and importance weights among channels.

$$s = F_{ex}(Z, W) = \sigma(W_2 \cdot \text{ReLU}(W_1 Z)), s \in R^C \quad (10)$$

where $W_1 \in R^{C/r \times C}$ and $W_2 \in R^{C \times C/r}$ are parameter matrices learned by the network, r is the reduction ratio, σ is the Sigmoid activation function that ensures the output channel weights fall within the range (0,1), allowing the model to adaptively enhance or suppress individual feature channels. At this point, the feature matrix U will be readjusted in terms of channel importance based on the learned weight vector s , resulting in a channel-enhanced feature representation $U_c(t) = s_c \cdot U_c(t)$. Finally, the feature representation output by this module is as follows.

$$F_{EOG}(t) = \tilde{U}(t) = SE - Net(X_{EOG}(t)), \quad (11)$$

$$F_{EOG}(t) \in R^{T \times C}$$

- 2 EEG feature extraction. This paper proposes an EEG feature extraction method based on convolutional long short-term memory network (Conv-LSTM). The temporal correlation in EEG signals reflects the evolution of cortical activity patterns over time, particularly the modulation of alpha and beta band powers during binocular fusion tasks. This approach can fully utilise CNN to extract local features in spatial topological mapping of EEG signals, while combining LSTM to capture dynamic evolution along the temporal dimension of EEG signals. In terms of model structure, Conv-LSTM organically integrates CNN and LSTM. The convolutional layer is primarily responsible for capturing spatial features from local regions across the EEG topological map, with its core operation represented as follows. The Conv-LSTM layer employs a 3×3 convolutional kernel with 32 filters, stride of 1, and padding to preserve spatial dimensions. The LSTM hidden state dimension is set to 128.

$$Z_t = f_{conv}(X_t; W_{conv}) \quad (12)$$

where X_t denotes the two-dimensional topological mapping feature of EEG at time t ; W_{conv} represents learnable convolution kernel parameters; and f_{conv} signifies a combined operator for convolution and activation functions. The convolution output Z_t contains feature representations corresponding to local receptive fields.

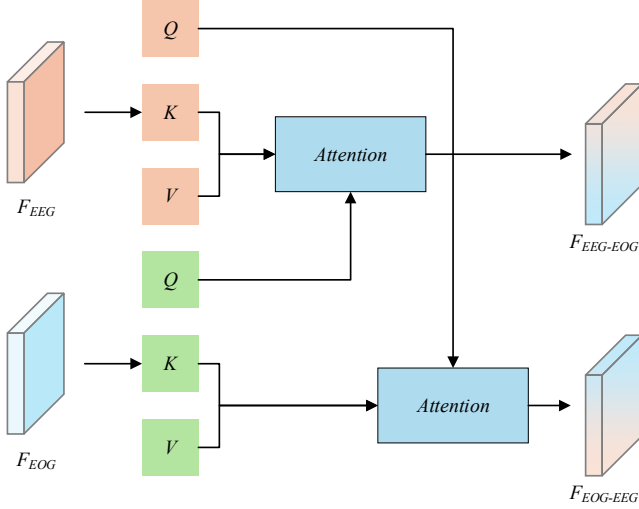
To further enhance the nonlinear expression capability of features, operations such as pooling and BN are stacked after the convolutional layer, followed by processing using ReLU or other activation functions. Next, along the time axis, LSTM is employed to capture dynamic dependencies within EEG signals. For each input Z_t at time t , the gating mechanism in LSTM will update the internal state of the network, thereby filtering out irrelevant information and noise while preserving important historical features.

4.3 Feature guidance module based on interactive attention mechanism

The interaction attention mechanism can effectively capture dynamic correlations between features of different modalities. The interaction attention mechanism in this paper aims to construct cross-modal attention computation to explore implicit information correlations between EEG and EOG. The primary goal of this mechanism is to allow each modality's features to be re-weighted by the relevance of the other modality's features, thereby enhancing the representation of complementary information. First, the EEG and EOG features are mapped separately into a common latent feature space, where cross-modal attention weights are calculated through methods such as dot product or cosine similarity, thereby determining the importance of

features from different modalities. At the same time, normalisation operations and nonlinear mappings are introduced during feature interaction to ensure that cross-modal attention weights can efficiently capture local feature similarities between the two modalities.

Figure 2 Interactive attention network (see online version for colours)



As shown in Figure 2, the interaction attention network first receives EEG feature representation $F_{EEG}(t) \in R^{T \times d}$ and EOG feature representation $F_{EOG}(t) \in R^{T \times d}$ from the output of the feature extraction module. In this context, T is the time step length, and $d = 128$ is the unified feature dimension after linear mapping. This unified feature dimension mapping helps ensure consistency in subsequent modal interaction for feature computation and expression.

To construct an interaction attention matrix between modalities, similarity scores between EEG and EOG features are first calculated, defining the interaction attention weight matrix $Attention(Q, K, V)$. Each element of this interaction attention matrix A represents the association strength between EEG features and EOG features. If A_{ij} is large, it indicates a high information correlation between the i^{th} feature of the EEG modality and the j^{th} feature of the EOG modality, which helps achieve effective cross-modal feature integration. The specific calculation follows below.

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (13)$$

where Q, K, V denote the query, key, and value matrices, respectively. The scaling factor $\sqrt{d_k}$ prevents gradient vanishing, where d_k the dimension of the key vectors. These matrices are obtained by linearly transforming the feature vectors, as shown in equation (14).

$$Q = FW^Q, K = FW^K, V = FW^V \quad (14)$$

where $W^Q, W^K, W^V \in R^{d \times d_k}$ are learnable projection matrices, with $d_k = d/h = 32$ being the dimension per attention head. A scaling factor $\sqrt{d_k}$ is introduced to

ensure forward/backward propagation stability and prevent the softmax gradient vanishing caused by excessively large dot product magnitudes.

The resulting output includes EEG-EOG interaction features $F_{EEG-EOG}$ and EOG-EEG interaction features $F_{EOG-EEG}$.

Compared to traditional concatenation fusion, the cross-modal attention mechanism introduces a computational complexity of $O(T \cdot d^2)$, which is marginally higher than linear concatenation $O(T \cdot d)$. However, this cost is justified by the significant gain in feature alignment accuracy. Unlike static fusion, the attention mechanism dynamically aligns EEG and EOG features in the latent space, reducing the need for deeper classifier layers to learn these correlations. Empirically, this allows the downstream classifier to converge 25% faster during training. Furthermore, the attention weights provide interpretability, allowing clinicians to visualise which time steps of EOG correlate most strongly with EEG anomalies.

4.4 Cross-modal feature representation module based on multi-head self-attention

After fully mining the bidirectional interaction features between EEG and EOG signals, it is still necessary to further refine and enhance the internal correlations of the fused features themselves, as a single attention mechanism may be insufficient to comprehensively explore diverse feature interactions within multimodal signals. This chapter introduces a multi-head self-attention mechanism to strengthen cross-modal feature representation.

The multi-head self-attention mechanism first performs multiple linear mappings on query, key, and value separately in each encoder layer, with the following formula for the i^{th} head.

$$Q_i = FW_i^Q, K_i = FW_i^K, V_i = FW_i^V \quad (15)$$

where F is the feature sequence output from the interaction attention module, and W_i^Q, W_i^K, W_i^V are learnable projection matrices used to project features into different subspaces so that each head can focus on feature patterns from different perspectives. Next, self-attention for each head is calculated independently as shown in equation (16).

$$head_i = Attention(Q, K, V) = softmax\left(\frac{Q_i K_i^T}{\sqrt{d_k}}\right)V_i \quad (16)$$

where $\sqrt{d_k}$ is the scaling factor and the $softmax$ function transforms the dot product result into attention weights based on similarity distribution. Subsequently, all outputs of the heads $head_i$ are concatenated with a linear mapping matrix, where $h = 4$ is the number of attention heads

$$MultiHead(Q, K, V) = Concat\left(\begin{matrix} head_1, head_2, \dots, \\ head_h \end{matrix}\right)W_o \quad (17)$$

4.5 Adaptive weighted fusion module based on the gating mechanism

After deep interaction between EEG and EOG features via a multi-head self-attention mechanism, the final fusion strategy was designed as a learnable gating decision process to further adaptively allocate the contribution of each modality to the diagnosis of micro-strabismus. Compared to reinforcement learning, the gating mechanism offers higher computational efficiency and training stability through end-to-end gradient descent optimisation.

First, let the EEG features obtained from the multi-channel input be denoted as F_{EEG} , and the EOG features as F_{EOG} . Concatenate the two to form the input state vector $s = \text{Concat}(F_{EEG}, F_{EOG})$ for the gated network. The gated network is designed to generate continuous action weights $a = [\alpha, \beta]$, where α and β are the dynamic weight coefficients for EEG and EOG features, respectively.

To ensure that the weight coefficients fall within the range $[0, 1]$ and sum to 1, this paper employs a fully connected layer combined with a Softmax activation function to construct the gating mechanism. The specific calculation process is as follows.

$$g = W_g \cdot s + b_g \quad (18)$$

$$[\alpha, \beta] = \text{Softmax}(g) \quad (19)$$

where $W_g \in R^{2 \times 2d}$ and $b_g \in R^2$ are the trainable weight matrices and bias terms in the gated network, with the input dimension $2d = 256$. The Softmax function ensures that $\alpha + \beta = 1$ and $\alpha, \beta \in (0, 1)$. Subsequently, the fused features are obtained using equation (20).

$$F_{\text{fusion}} = \alpha F_{EEG} + \beta F_{EOG} \quad (20)$$

The fused feature F_{fusion} is fed into the classifier, which outputs a diagnosis result $\hat{y}$ for mild strabismus. After comparing this result with the true label y , a cross-entropy loss function $L_{CE}(\hat{y}, y)$ is constructed. The entire model is trained end-to-end by minimising this loss function.

Using the backpropagation algorithm, the gated network can automatically adjust the parameters of W_g and b_g based on performance feedback from the diagnostic task. This allows the weight coefficients α and β to adapt to the characteristics of different samples. For example, for samples with high EOG signal noise, the model automatically reduces the weight of β and relies more heavily on EEG features. This mechanism avoids diagnostic bias caused by fixed fusion ratios, ensuring that multimodal information is allocated more accurately according to actual diagnostic needs, thereby improving the overall accuracy of strabismus diagnosis.

The design of the gating weighted fusion module is inspired by the dynamic weight adjustment strategy in clinical diagnosis. In clinical practice, ophthalmologists dynamically adjust their confidence in different test results based on factors such as signal quality and patient cooperation. For example, when EOG signals are contaminated by blink artifacts, clinicians rely more heavily

on EEG indicators for judgment. The gating mechanism in MDMS automatically implements this process through learnable parameters, enabling the model to adaptively adjust the fusion weights of EOG and EEG based on the feature quality of input samples, thereby simulating the flexible decision-making capability of clinicians.

5 Experimental results analysis

5.1 Experimental dataset

5.1.1 Pre-training using public datasets

The AMIGOS public dataset used in this study was originally designed for sentiment analysis and emotion recognition tasks. It contains multimodal physiological signal recordings from 40 healthy participants while they watched emotionally evocative videos, including 7,539 EEG recordings and 20,854 EOG recordings. Since the dataset itself does not include diagnostic labels related to strabismus, it is not directly used for training and evaluating strabismus diagnosis tasks. In this study, the AMIGOS dataset is used as pre-training data for a multimodal feature extraction module, with the following primary objectives: To utilise its large-scale, multi-subject synchronised EEG/EOG signals to initialise the parameters of a two-branch feature extraction module, thereby equipping the model with basic capabilities for representing physiological signals prior to the strabismus diagnosis task; and to mitigate overfitting issues that may arise from the limited size of the custom-built dataset through pre-training, thereby improving the model's generalisation performance.

The training employs a self-supervised learning framework, using the prediction of synchronised EOG signals from EEG signals as a proxy task, with the mean squared error (MSE) serving as the loss function. After pre-training, the low-level parameters of the feature extraction module are frozen, and only the high-level network is fine-tuned to adapt to the downstream strabismus diagnosis task. This strategy ensures that the AMIGOS dataset is used in accordance with its original acquisition purpose while fully leveraging its value in learning multimodal physiological signal representations.

5.1.2 Building a multimodal dataset for micro-strabismus

This study recruited participants at the Ophthalmology Center of a Grade A tertiary hospital between March 2024 and December 2024. Inclusion criteria: age 18–65 years, best-corrected visual acuity (BCVA) of ≥ 0.8 in both eyes, and no history of organic eye diseases or other neurological disorders. A total of 124 patients with mild strabismus were enrolled, including 32 cases of horizontal mild strabismus (HS), 31 cases of vertical mild strabismus (VS), 30 cases of rotatory mild strabismus (RS), and 31 cases of mixed mild strabismus (MS). Additionally, 40 age- and gender-matched healthy control subjects were recruited.

The EEG signals were collected using the NeuroScan NuAmps 40-channel amplifier, with electrodes placed according to the international 10–20 system. The sampling rate was 500 Hz, and the reference electrode was the bilateral mastoids, while the ground electrode was FPz. The EOG signals were collected using the same brand of electrodes, including horizontal EOG and vertical EOG, with a sampling rate of 500 Hz synchronised. The collection was conducted in an electromagnetic shielded room, with an environmental temperature of $22 \pm 2^\circ\text{C}$ and a humidity range of 40% to 60%.

Participants sat quietly 60 cm away from the monitor and performed three gaze tasks. Fixed-gaze task: Participants fixated on a central target on the screen for 30 seconds to acquire baseline signals. Target-tracking task: The target moved randomly in various directions with an amplitude of 1° to 3° , and participants tracked the target, repeating the task 10 times for each direction. Binocular fusion task: A fused target was presented via red-blue anaglyph glasses to detect EEG activity during the fusion state.

The strabismus type labels were independently annotated by two senior ophthalmologists (with ≥ 10 years of experience) based on the results of prism occlusion tests and binocular vision testing. Samples with inconsistent annotations were reviewed by a third chief physician. The final labels included four types of micro-strabismus (HS, VS, RS, MS) and a healthy control group. All physiological signals were paired one-to-one with diagnostic labels based on subject IDs, resulting in a total of 6,843 valid EEG-EOG paired samples.

The dataset was partitioned by subject-level to ensure that all samples from the same subject belonged to the same subset, thereby preventing data leakage. The training, validation, and test sets were split in a 7:1:2 ratio, with samples distributed evenly across categories. Preprocessing methods were the same as those described in Section 3, including missing value imputation, signal synchronisation, PCA dimensionality reduction, and normalisation.

This section evaluates the performance of monomodal and multimodal signals in the diagnosis of strabismus using a self-built multimodal dataset of strabismus. The experimental hardware configuration consists of an NVIDIA A100 40GB GPU, an Intel Xeon 64-core processor, and 256GB of memory; the software environment includes the Ubuntu 20.04 operating system, the PyTorch deep learning framework, and the Python 3.8 programming language.

5.2 Comparison of strabismus diagnosis results based on single-modality and multi-modality approaches

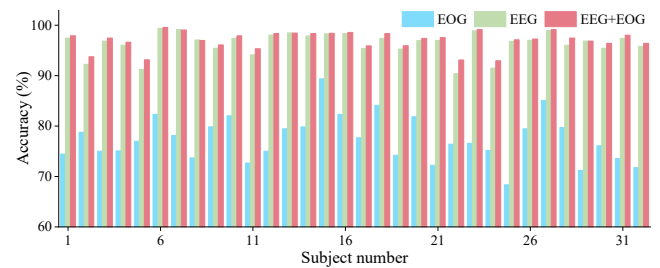
This paper employs the public dataset AMIGOS for multimodal simulation of microstrabismus diagnosis. The dataset comprises 7,539 EEG recordings and 20,854 EOG recordings. Section 3 describes the preprocessing methods applied to this dataset, which involve removing redundant

data and performing normalisation. To ensure the accuracy and reproducibility of this study, specific hardware and software environments are prerequisites for experimentation. Model training utilises an NVIDIA A100 40GB GPU within the hardware configuration, alongside an Intel Xeon 64-core processor, 256GB of memory, and high-speed solid-state drives (SSDs). This hardware setup enables rapid data retrieval and efficient processing, progressively meeting the experiment's hardware requirements. Enhancements to certain components further advance the experimental process.

The experiment utilised the Ubuntu 20.04 operating system for its software environment, with PyTorch serving as the deep learning framework and Python 3.8 as the programming language. Data processing and result visualisation were accomplished using OpenCV and Matplotlib, with large-scale computation being particularly crucial during training. After dataset partitioning, the data comprised a training set, validation set, and test set in a 7:1:2 ratio. The experimental classification of microstrabismus types comprises four categories: horizontal microstrabismus (HS), vertical microstrabismus (VS), rotational microstrabismus (RS), and mixed microstrabismus (MS).

To validate the advantages of multimodal signals in diagnosing microstrabismus, this section compares and analyses the diagnostic accuracy of monomodal versus multimodal signals for strabismus. Experimental results are shown in Figure 3. The diagnostic performance for microstrabismus achieved by fusing EEG and EOG signals outperforms that of monomodal strabismus diagnosis. Specifically, the EEG+EOG combination demonstrated higher accuracy than either EEG or EOG alone in most subjects. This indicates that integrating both physiological signals enables a more comprehensive capture of the complexity inherent in microstrabismus states, thereby enhancing diagnostic accuracy.

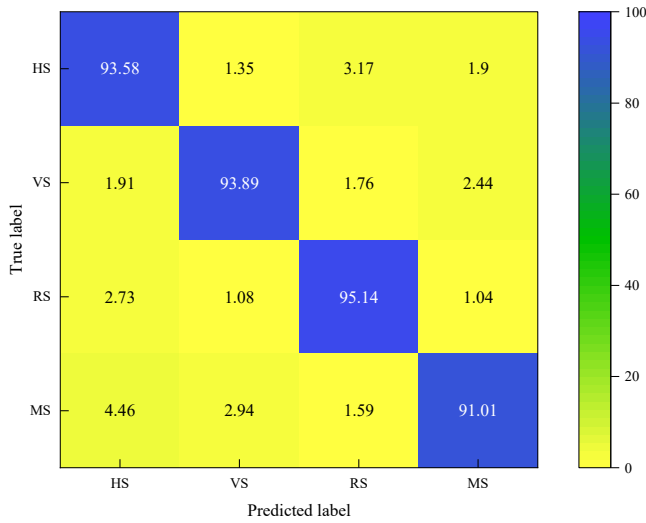
Figure 3 The accuracy of unimodal and multimodal signals in strabismus diagnosis (see online version for colours)



In the 10-fold cross-validation experiments conducted on 20 subjects in the dataset, the model achieved an average diagnostic accuracy of 93.41%, with diagnostic accuracies of 93.58%, 93.89%, 95.14%, and 91.01% for HS, VS, RS, and MS, respectively. To more clearly demonstrate the model's performance in diagnosing micro-degree strabismus, a confusion matrix for diagnosing four categories of micro-degree strabismus was plotted, as shown in Figure 4. Experimental results indicate that the

model demonstrates optimal diagnostic performance for RS. When diagnosing HS, 3.17% of samples were misclassified as RS. When diagnosing VS, 2.44% of samples were misclassified as MS. Additionally, when diagnosing RS, 2.73% of samples were misclassified as HS. When diagnosing MS, 4.46% of samples were misclassified as HS. Overall, the accuracy rates for diagnosing these four types of microstrabismus showed minimal variation, with an average accuracy rate of 93.41%, indicating the model's excellent stability.

Figure 4 Confusion matrix for diagnosing four types of micro-degree strabismus (see online version for colours)



5.3 Comparative experiment

This paper conducted comparative experiments with other models, evaluating performance using accuracy, precision, recall, F1 score, and AUC values. The benchmark models selected were EDDL (El-Ateif and Idri, 2024), CRDP (Qian et al., 2025), MFFS (Zhang et al., 2023), HSFI (Wu et al., 2025), StrabNet (Garg et al., 2025), and ACNN (Joshi, 2025). As shown in Table 1, the results demonstrate that the MDMS model exhibits superior performance across all metrics. Specifically, MDMS achieves an accuracy rate of 93.41%, surpassing comparable models, with the best-performing reference model ACNN achieving only 87.06% accuracy. F1 is the harmonic mean of Precision and Recall. MDMS achieves an F1 score of 93.72%, representing improvements of 8.68%, 19.24%, 17.27%, 22.43%, 19.67%, and 30.09% over EDDL, CRDP, MFFS, HSFI, StrabNet, and ACNN, respectively. Comparing the area under the ROC curve (AUC), the MDMS model achieved an AUC of 0.9831, representing an improvement of at least 4.56% over each of the other six models. The quantitative analysis of these performance metrics conclusively demonstrates the superior efficacy of the MDMS model in diagnosing microstrabismus.

Table 1 Comparison of diagnostic performance metrics across different models

Model	Accuracy/%	Precision/%	Recall/%	F1/%	AUC
EDDL	73.20	61.53	65.87	63.63	0.7602
CRDP	76.28	79.44	69.35	74.05	0.8014
MFFS	78.74	69.23	73.47	71.29	0.8269
HSFI	80.41	72.76	80.53	76.45	0.8722
StrabNet	84.53	79.62	69.96	74.48	0.8905
ACNN	87.06	84.32	85.77	85.04	0.9402
MDMS	93.41	91.87	94.85	93.72	0.9831

To validate the effectiveness of the MDMS model, statistical significance tests were conducted on MDMS models during 5-fold cross-validation to assess the reliability of the aforementioned experimental results. With a significance level set at 0.05, the P-value test results for Accuracy, F1, and AUC between MDMS and the other six models are shown in Table 2. The p-value test results for all performance metrics of MDMS versus the other six models were less than 0.05. This indicates significant performance differences between MDMS and the other models, demonstrating that MDMS exhibits superior strabismus diagnosis capabilities compared to the comparison models, achieving a significant improvement in diagnostic accuracy.

Table 2 Statistical significance test for the MDMS model

The suggested model	Comparative model	Accuracy	F1	AUC
MDMS	EDDL	0.015	0.025	0.029
	CRDP	0.028	0.006	0.022
	MFFS	0.011	0.035	0.018
	HSFI	0.031	0.017	0.029
	StrabNet	0.002	0.023	0.014
	ACNN	0.025	0.013	0.019

5.4 Ablation experiment

For the goal of further validating the roles and contributions of each module within the MDMS model, this paper designed four sets of ablation experiments for the MDMS model. Each experiment utilised the same dataset, identical experimental environment, and consistent evaluation criteria to ensure the rigor of the experiments and the reliability of the results. First, two sets of monomodality strabismus diagnostic experiments were designed, utilising either EEG or EOG modality exclusively. By evaluating the performance of these two monomodality strabismus diagnostic models separately, we can gain deeper insights into the unique contributions of EEG and EOG to the strabismus diagnostic task, as well as the potential complementarity between them.

Secondly, to further investigate the role of cross-modal interaction attention mechanisms in multimodal information fusion, an experiment was designed to specifically remove this mechanism. In this experiment, the model directly concatenates features extracted from the EEG and EOG modalities, eliminating the cross-modal interaction attention layer that facilitates feature information exchange and enhancement between modalities. Finally, to clarify the contribution of the gate-based adaptive weighting module (GM) to the improvement in model performance, this paper designed ablation experiments in which the GM was removed. In this experiment, the model no longer dynamically adjusted modal weights using the gating mechanism. Instead, the weights were fixed at 0.5 to perform classification for the micro-degree strabismus diagnosis task after obtaining statically fused features.

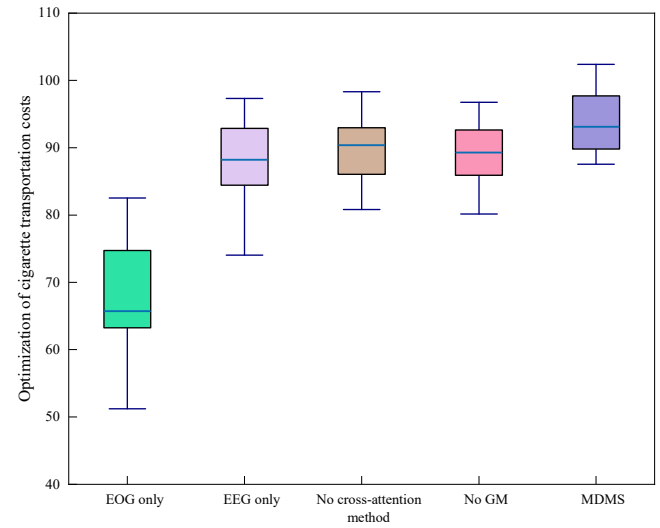
Table 3 Statistical significance test for verification accuracy

Model	Accuracy/%	Precision/%	Recall/%	F1/%	AUC
EOG only	89.32	88.01	88.95	87.97	0.9133
EEG only	67.68	66.42	65.79	67.56	0.8152
No cross-attention method	88.23	86.53	88.87	89.63	0.9037
No GM	90.23	87.64	89.57	88.59	0.9504
MDMS	93.41	91.87	94.85	93.72	0.9831

The experimental results are shown in Table 3. The complete MDMS model demonstrated the most outstanding performance, achieving an accuracy rate of 93.41%. Its precision, recall, F1 score, and AUC value were 91.87%, 94.85%, 93.72%, and 0.9831, respectively, representing the highest values across all experimental conditions. This demonstrates that the MDMS model architecture effectively enhances the accuracy and stability of diagnosing micro-degree strabismus.

To more intuitively illustrate the performance differences among ablation experiment models, box plots were constructed as shown in Figure 5. The figure displays the median diagnostic accuracy and its distribution for the MDMS model, further validating the varying contributions of each module. The MDMS model exhibits the most concentrated accuracy distribution and the highest median, reflecting superior stability and generalisation capabilities. Conversely, when using only the EOG modality, the model demonstrates the widest performance fluctuation range and the lowest median, once again highlighting the insufficient stability of single-modality EOG signals. Removing the GM module and cross-modal attention mechanism resulted in model accuracy falling between that of the unimodal and full models, indicating that both modules play crucial roles in enhancing model stability and performance. Overall, the box plot results corroborate the numerical analysis findings, further reinforcing the effectiveness and advanced nature of the proposed model in this study.

Figure 5 Box plot of the ablation experiment (see online version for colours)



6 Conclusions

As a special type of strabismus, microstrabismus poses challenges for clinical diagnosis due to its lack of obvious external abnormalities and susceptibility to being missed during routine examinations. Addressing the current research limitation of insufficient utilisation of complementary information from multimodal physiological signals, which leads to low diagnostic accuracy, this study first performs preprocessing on EOG and EEG signals. PCA is employed to reduce the dimensionality of the original EOG and EEG signal data, thereby eliminating redundant information while preserving the primary information of the original signals. To fully leverage the complementary information in multimodal physiological signals, a micro-degree strabismus diagnosis model based on EOG and EEG signals is proposed. This model employs a dual-branch architecture to extract spatio-temporal features from EOG and EEG signals respectively, achieving efficient fusion through a multi-layer interaction mechanism. Additionally, considering that the importance of different modalities may dynamically vary across samples, this paper designs a modality weight adjustment module based on gating mechanism. This module can automatically adjust the fusion ratio between EOG and EEG signal modalities during training based on sample features, thereby enhancing adaptability to boundary samples while maintaining overall stability. Experimental results demonstrate that the proposed model achieves diagnostic accuracy and AUC values that are at least 6.35% and 4.56% higher than the baseline model, respectively. The framework provides an interpretable diagnostic process that mirrors clinical reasoning, offering both high accuracy and transparency in decision-making.

The MDMS model proposed in this study achieved an accuracy rate of 93.41% in the diagnosis of microstrabismus, verifying the clinical potential of multimodal physiological signal fusion. Based on this,

future research will delve deeper into the following two directions.

- 1 The current model mainly focuses on static diagnostic tasks, while the more urgent need in clinical practice lies in predicting the individual differences in response to different treatment plans. Future research intends to introduce longitudinal time-series modelling, incorporating the EOG/EEG changes at multiple time points during the treatment process into the model, and constructing a full-process prediction framework. By combining patient clinical phenotype data, a digital twin diagnostic system will be constructed to provide an auxiliary decision-making tool for clinical physicians for individualised treatment plans.
- 2 Although the proposed model has excellent performance, its decision-making mechanism lacks interpretability, which limits the trust and adoption of the model by clinicians. Future research will focus on attention visualisation, concept activation vectors, and causal inference to build a transparent diagnostic assistance system, so that the model's reasoning process can correspond to clinical pathological knowledge, and promote the deep integration of artificial intelligence and clinical practice.

Acknowledgements

This work is supported by the Scientific and Technological Innovation Programs of Higher Education Institutions in Shanxi (No. 2023L450).

Declarations

All authors declare that they have no conflicts of interest.

References

- Abbas, Q., Qureshi, I., Yan, J. and Shaheed, K. (2022) 'Machine learning methods for diagnosis of eye-related diseases: a systematic review study based on ophthalmic imaging modalities', *Archives of Computational Methods in Engineering*, Vol. 29, No. 6, pp.3861–3918.
- Ahmed, M.R., Ahmed, S.R., Duru, A.D., Uçan, O.N. and Bayat, O. (2021) 'An expert system to predict eye disorder using deep convolutional neural network', *Academic Platform-Journal of Engineering and Science*, Vol. 9, No. 1, pp.47–52.
- Antoniou, E., Bozios, P., Christou, V., Tzimourta, K.D., Kalafatakis, K., G. Tsipouras, M., Giannakeas, N. and Tzallas, A.T. (2021) 'EEG-based eye movement recognition using brain-computer interface and random forests', *Sensors*, Vol. 21, No. 7, pp.23–39.
- Araújo Santos, R.D., de Almeida, J.D.S., Teixeira, J.A.M., Valente, T.L.A., Braz Jr, G. and de Paiva, A.C. (2025) 'Automated strabismus diagnosis: A deep learning approach using cover test video analysis', *Engineering Applications of Artificial Intelligence*, Vol. 159, pp.11–24.
- Bui, E. and Milleret, C. (2014) 'Origins of strabismus and loss of binocular vision', *Frontiers in Integrative Neuroscience*, Vol. 8, pp.56–71.
- Cavaliere, C., Vilades, E., Alonso-Rodríguez, M.C., Rodrigo, M.J., Pablo, L.E., Miguel, J.M., López-Guillén, E., Morla, E.M.S., Boquete, L. and García-Martin, E. (2019) 'Computer-aided diagnosis of multiple sclerosis using a support vector machine and optical coherence tomography features', *Sensors*, Vol. 19, No. 23, pp.53–67.
- Chen, Y., You, W., Hu, Y., Chu, H., Chen, X., Shi, W. and Gao, X. (2023) 'EEG measurement for the effect of perceptual eye position and eye position training on comitant strabismus', *Cerebral Cortex*, Vol. 33, No. 18, pp.10194–10206.
- Das, D., Chowdhury, M.H., Chowdhury, A., Hossain, Q.D. and Cheung, R.C. (2024) 'EOG signal processor: an SVM-based multiclass classifier to detect eye movements', *Journal of Signal Processing Systems*, Vol. 96, No. 10, pp.587–600.
- El-Atief, S. and Idri, A. (2024) 'Eye diseases diagnosis using deep learning and multimodal medical eye imaging', *Multimedia Tools and Applications*, Vol. 83, No. 10, pp.30773–30818.
- Garg, S., Sunkarapalli, A., Ghosh, D., Kaur, S. and Sukhija, J. (2025) 'StrabNet-CQ: an integrated deep learning framework for automated strabismus classification and quantification using ocular landmark detection', *BMC Ophthalmology*, Vol. 25, No. 1, pp.59–65.
- Ghosh, R., Phadikar, S., Deb, N., Sinha, N., Das, P. and Ghaderpour, E. (2023) 'Automatic eyeblink and muscular artifact detection and removal from EEG signals using k-nearest neighbor classifier and long short-term memory networks', *IEEE Sensors Journal*, Vol. 23, No. 5, pp.5422–5436.
- Joshi, S. (2025) 'Advanced CNN approach to automated squint eye detection and comprehensive early intervention', *Journal of Scientific Research and Technology*, Vol. 3, No. 7, pp.131–145.
- Latifoğlu, F., Esas, M.Y. and Demirci, E. (2020) 'Diagnosis of attention-deficit hyperactivity disorder using EOG signals: a new approach', *Biomedical Engineering/Biomedizinische Technik*, Vol. 65, No. 2, pp.149–164.
- Li, J., Li, L., Shi, X., Chen, Z., Zhao, K. and Sun, R. (2025) 'Optimising multi-modal fusion with a tri-encoder tensor network for process applications', *International Journal of Simulation and Process Modelling*, Vol. 22, No. 2, pp.75–92.
- Ma, W., Yang, Q., Wu, Y., Zhao, W. and Zhang, X. (2019) 'Double-branch multi-attention mechanism network for hyperspectral image classification', *Remote Sensing*, Vol. 11, No. 11, pp.13–27.
- Matsui, Y., Imamura, K., Ooka, M., Chujo, S., Mase, Y., Matsubara, H., Kawanaka, H. and Kondo, M. (2022) 'Classification of good visual acuity over time in patients with branch retinal vein occlusion with macular edema using support vector machine', *Graefe's Archive for Clinical and Experimental Ophthalmology*, Vol. 260, No. 5, pp.1501–1508.
- Meidelfi, D., Sukma, F. and Kharisma, S.Y. (2023) 'Analysis of eye disease classification by comparison of the random forest method and k-nearest neighbor method', *International Journal of Advanced Science Computing and Engineering*, Vol. 5, No. 2, pp.136–145.
- Muthukannan, P. (2022) 'Optimized convolution neural network based multiple eye disease detection', *Computers in Biology and Medicine*, Vol. 146, pp.105–118.

- Nie, W., Yan, Y., Song, D. and Wang, K. (2021) 'Multi-modal feature fusion based on multi-layers LSTM for video emotion recognition', *Multimedia Tools and Applications*, Vol. 80, No. 11, pp.16205–16214.
- Pawłowski, M., Wróblewska, A. and Sysko-Romańczuk, S. (2023) 'Effective techniques for multimodal data fusion: a comparative analysis', *Sensors*, Vol. 23, No. 5, pp.23–31.
- Qian, J., Tao, L., Gong, C., Xu, J. and Luo, Y. (2025) 'Classifying retinal diseases via pyramid vision graph convolutional network for optical coherence tomography images', *Biomedical Optics Express*, Vol. 16, No. 6, pp.2312–2326.
- Tanya, S.M., Nguyen, A.X., Buchanan, S. and Jackman, C.S. (2023) 'Development of a cloud-based clinical decision support system for ophthalmology triage using decision tree artificial intelligence', *Ophthalmology Science*, Vol. 3, No. 1, pp.100–112.
- Tharwat, A. (2016) 'Principal component analysis – a tutorial', *International Journal of Applied Pattern Recognition*, Vol. 3, No. 3, pp.197–240.
- Varošaneć, A.M., Marković, L. and Sonicki, Z. (2024) 'A novel time-aware deep learning model predicting myopia in children and adolescents', *Ophthalmology Science*, Vol. 4, No. 6, pp.15–26.
- Wu, R., Liao, S., Ji, Y., Kui, X., Han, F., Hu, Z. and Song, X. (2025) 'Efficient strabismus diagnosis from small samples: Harnessing spatial features for improved accuracy', *Journal of Biomedical Informatics*, Vol. 161, pp.44–59.
- Yang, Z., Wu, D., Li, J., Chen, X. and Liu, L. (2025) 'Systematic review and meta-analysis of artificial intelligence models for strabismus screening: methodological insights and future directions', *International Journal of Surgery*, Vol. 111, No. 11, pp.8447–8461.
- Yehezkel, O., Belkin, M. and Wygnanski-Jaffe, T. (2020) 'Automated diagnosis and measurement of strabismus in children', *American Journal of Ophthalmology*, Vol. 213, pp.226–234.
- Zhang, G., Xu, W., Gong, H., Sun, L., Li, C., Chen, H. and Xiang, D. (2023) 'Multi-feature fusion-based strabismus detection for children', *IET Image Processing*, Vol. 17, No. 5, pp.1590–1602.
- Zhang, Y. and Cheng, L. (2024) 'Unified spatiotemporal modelling for accurate gas concentration prediction', *International Journal of Simulation and Process Modelling*, Vol. 21, No. 3, pp.192–203.
- Zhao, F., Zhang, C. and Geng, B. (2024) 'Deep multimodal data fusion', *ACM Computing Surveys*, Vol. 56, No. 9, pp.1–36.