Z. Zeng, L. Tao, J. Hu, R. Su, L. Meng, C. Chen, and W. Chen, “Multi-scale Inception-based Deep Fusion Network for electrooculogram-based eye movements classification,” Biomedical Signal Processing and Control, vol. 103, pp. 107377, 2025/05/01/, 2025.
DOI: 10.1016/j.bspc.2024.107377

# Multi-scale Inception-based Deep Fusion Network for Electrooculogram-based Eye Movements Classification

Zheng Zeng[1], Linkai Tao[1], Jun Hu, Ruizhi Su, Long Meng, Chen Chen*, and Wei Chen* *Member, IEEE*

***Abstract*—Classifying eye movements accurately is essential for various practical applications, such as human-computer interfaces, sleep staging, and fatigue detection. However, eye movement classification (EMC) based on electrooculogram (EOG) is still challenging, and the existing solutions are still suboptimal in terms of accuracy. Traditional machine learning (ML)-based methods mainly focus on hand-crafted features, relying heavily on prior knowledge of EOG analysis. Besides, most existing deep learning (DL)-based methods simply concentrate on extracting sing-scale or multi-scale features without considering the contribution of features across different levels, constraining the model capacity in learning discriminative representations. To address the aforementioned problems, a novel Multi-scale Inception-based Deep Fusion Network (MIDF-NET), composed of paralleled CNN streams and a multi-scale feature fusion (MSFF) module, is proposed to extract informative features from raw EOG signals. The paralleled CNN streams can extract multi-scale representations of EOGs effectively and the MSFF module fuses these features, taking advantage of low and high-level multi-scale features. Comprehensive experiments were conducted on 5 public EOG datasets (50 subjects and 59 recordings), containing 5 types of eye movements (Blink, Up, Down, Right, and Left). State-of-the-art EOG-based eye movement approaches including classical machine learning models and deep networks were also implemented for comparison. Experimental results demonstrate that our MIDF-NET achieved the highest accuracy among the 5 public datasets (87.7%, 86.0%, 95.0%, 94.2%, and 95.4%), outperforming state-of-the-art methods with a significant accuracy improvement. In conclusion, the proposed MIDF-NET can comprehensively consider the multi-level features according to the feature fusion sub-networks and effectively classify the eye movement patterns via the enhanced representation of EOGs.**

***Index Terms*—Electrooculogram (EOG), Eye movement classification (EMC), Deep learning;**

## I. Introduction

The eyes play a crucial role in the human body. In human-computer interaction (HCI) research, the eyes serve as both receptors of visual information and emitters for interaction through eye movement. Therefore, utilizing eye movements as emitters in HCI offers direct, simple, and efficient advantages compared to other types of emitters, such as limb movements, especially for tasks that heavily rely on visual feedback. Eye movement classification (EMC) is a commonly used technique in EOG-HCI to translate eye movements into interaction commands. There have been many research achievements in different EOG-based EMC applications [1]–[4], etc. Compared to prevailing eye-tracking techniques (i.e., image/video-based techniques and search coils), EOG signals are less susceptible to environmental influences. Besides, EOG signals have low user cooperation requirements and relatively low data acquisition costs [5]. With the rapid development of flexible electronics and wearable sensor technologies, decoding user eye movement patterns using EOG signals is becoming increasingly feasible, providing a convenient biomedical sensing solution.

However, most existing studies in the field still primarily explore eye movement recognition technology within specific interactive application domains. The advantage of such strategies lies in the additional conditions associated with specific interactive tasks. The developed eye movement recognition algorithms do not necessarily need to focus on the characteristics of eye movements to achieve relatively objective eye movement recognition performance in that scenario. However, the results of such studies cannot be effectively transferred to other research within this field, ultimately leading to a lack of accumulation of knowledge in the EOG-HCI domain.

Existing EOG-based EMC approaches can be broadly categorized into traditional approaches, machine learning approaches, and deep learning approaches. For reliable EMC performance, EOG-based EMC approaches require extracting the discriminative EOG features to fully represent diverse eye movements. Traditional approaches (i.e., threshold-based approach and waveform matching [6]–[8]) mainly focus on the low-level features with sufficient detailed information (i.e., waveform shape, amplitude, waveform duration). These low-level features can characterize the shape information to distinguish eye movements since different eye movements often show EOG morphology differences in EOG signals. However, the low-level features are insufficient when classifying more eye movements. The low-level features have rich detailed information but also contain more noise and lack semantic information. Therefore, the performance of traditional approaches is unstable, requiring repetitive calibrations.

The machine learning (ML) approach is another prevailing alternative for EOG-based EMC. In contrast to the traditional approach, the ML approach integrates more sufficient hand-crafted features through feature engineering, thereby enhancing the distinguishability of eye movement patterns. Various hand-crafted features (i.e., multi-scale features, temporal features, spectral features, etc.) have been explored [8]–[12]. Especially, the multi-scale features (Fourier transforms, Empirical mode decompositions, Wavelet decomposition, etc.) have been proven to provide discriminative EOG representations of diverse eye movement [13]. However, the challenge with ML approaches is that the ML features rely heavily on prior knowledge. Targeting different electrode arrangements, the ML features also require to be designed laboriously. In addition, the extraction of hand-crafted features is manual and time-consuming. The limited feature engineering capability may implicitly cause information loss, leading to a suboptimal hand-crafted feature map.

Recent deep learning (DL) approaches have shown superior performance in EOG-based EMC. Some general DL architectures, such as convolutional neural networks (CNN), recurrent neural networks (RNN), and long short-term memory (LSTM), are employed to extract informative features from raw EOG signals. The DL approaches can automatically learn the deep features to infer diverse eye movement patterns. In comparison with the low-level features or ML features, these deep features, after multiple convolution operations, may have stronger semantic information to well distinguish eye movements. Existing research has shown that satisfactory performance can already be achieved when using simple CNN architectures. However, the deep model has not been thoroughly investigated yet to fully extract the discriminative features. More exquisite architectures can be specifically designed for EOG-based EMC tasks. For instance, the simple CNN architectures can only generate the fixed scale features, making it difficult to capture the multi-scale information in EOG signals. Besides, the deep features are often high-level and lack details after multi-layer convolutions. As the detailed information can reflect the EOG morphology differences, it is crucial to fully utilize the low-level and high-level features to well represent different eye movements. By properly fusing deep features from different levels, the classifiers can synchronously benefit from detail and semantic information, thereby further boosting performance.

Based on the above concerns, we propose a novel Multi-scale Inception-based Deep Fusion Network (MIDF-NET) to improve the accuracy of EOG-based EMC. The MIDF-NET is hierarchical and can learn the discriminative deep features automatically. The backbone of the MIDF-NET efficiently integrates two self-designed Inception modules to encode the raw EOG signals into multi-scale representations, offering MIDF-NET superior capability in capturing the scale information during eye movements. In addition, to take full advantage of detail and semantic information, a multi-scale feature fusion (MSFF) module is proposed to fuse the multi-level EOG features. The MSFF module equally considers the importance of deep features from different levels. In this regard, the MIDF-NET can synchronously benefit from the low-level and high-level features to infer diverse eye movements. We evaluate our model using the Leave-One-Subject-Out Cross-Validation strategy on multiple datasets (5 datasets, 50 subjects, and 59 recordings). Additionally, 8 state-of-the-art EOG-based eye movement methods (including 6 classical ML methods and 3 DL methods) were implemented on the same datasets for a direct and fair comparison. The main contribution of this work is two-fold:

1) MIDF-NET, a novel deep learning approach is proposed. The MIDF-NET offered multi-scale processing to extract the features from the raw EOG signals. Meanwhile, an MSFF module was integrated into the MIDF-NET backbone to consider the advantages of low and high-level features. The classification accuracy of the MIDF-NET is promising compared with the state-of-the-art methods, demonstrating its superiority in classifying multiple eye movements.

2) Our approach is robust enough and achieves promising performance under various conditions (multiple subjects, electrode schemes, and data acquisition scenes). The experimental results are comprehensive and may provide a public benchmark for EOG-based EMC.

This paper is organized as follows. In Section II, we introduce the 5 public datasets and data preprocessing. In Section III, the network architecture is described. The results are presented in Section IV. In Section V, we discuss the results and compare our proposed method with state-of-the-art methods. In the last section, we conclude the paper and present the limitations of this study.

## II. Related Work

This section presents an overview of previous studies on EOG-based EMC. We selected some representative works that used DL or ML methods for EOG-based EMC recognition from 2010 to the present. The presentation includes the technical foundation of the algorithms used in these works, validation methods, classification tasks, and sample sizes. We also present the corresponding aspects of our work (As shown in Table I). The relevant works of traditional methods are not included in the table since traditional approaches are unstable in classifying multiple eye movements.

ML-based approaches mainly employed hand-crafted features to represent different eye movements. In addition, various ML models including the support vector machine (SVM), artificial neural network (ANN), K-nearest neighbor (KNN), decision trees (DT), and ensemble models have been used to classify multiple eye movements. Andreas *et al.* utilized the temporal features to analyze the eye movement information [10]. Lv *et al.* extracted the spatial EOG features to infer the eye movements (Up, Down, Right, and Left) by linearly projecting pre-processed EOG signals to the spatial filter [14]. An accuracy of 99% on 10 subjects was reported by employing the SVM classifier. Lopez *et al.* proposed a two-stage ensemble model for EOG-based EMC (Up, Down, Right, Left, and Blink) [8]. The raw EOG signals were directly fed into the ensemble model. An accuracy of 99.9% was reported on 6 subjects. Recent studies have also demonstrated that the multi-scale analysis can provide discriminative representations for EOG-based EMC. Khan *et al.* employed the empirical wavelet transform to decompose the EOG signals [13]. By extracting the multi-scale features, a classification accuracy of 98.9% was achieved on 10 subjects. However, the ML approaches rely heavily on prior knowledge to incorporate the informative features. In addition, the EOG data are usually generated by heterogeneous sources (different subjects or electrode schemes), embracing sufficient differentiated information. Therefore, it is challenging to well capture the eye movement patterns with hand-crafted features. The ML approaches generally fail to automatically extract and employ the discriminative features describing diverse eye movements.

On the contrary, due to the large number of learnable parameters in models constructed using DL, the training of these models does not heavily rely on features proposed by prior experience. Therefore, in the analysis of more complex tasks, they often outperform ML. Recent DL solutions mainly leverage the one-dimensional or two-dimensional CNN, LSTM, or hybrid CNN-LSTM architecture for EOG-based EMC. Zou *et al.* [15] proposed a single-scale CNN network to decode eye movement patterns from 6 subjects via EOG signals. The recognition accuracy can be up to 90.47 %. Aya *et al.* [16] utilized the long-short term memory (LSTM) to classify the horizontal eye movements (right and left). The proposed LSTM networks captured the temporal and nonlinear structure of EOG signals and achieved an average accuracy of 90.1 %. Gu *et al.* first convoluted the raw EOG signals [17]. Subsequently, a layered RNN was employed to infer the eye movements. A classification accuracy of 90.72% was achieved on 20 subjects. However, existing DL approaches have not fully considered the physiological characteristics of EOGs. For instance, the single-scale CNN architectures use only one temporal scale, which may constrain the model's capacity to capture multi-scale information during eye movements. Besides, important detailed information on EOG signals is often lost in the process of multi-layer propagation, which is also crucial for accurate eye movement recognition. The existing DL models for EOG-based EMC are by means of simple deep architectures. Unfortunately, this represents a significant drawback since the extracted deep features may not fully depict the eye movement patterns. For EOG-based EMC tasks,

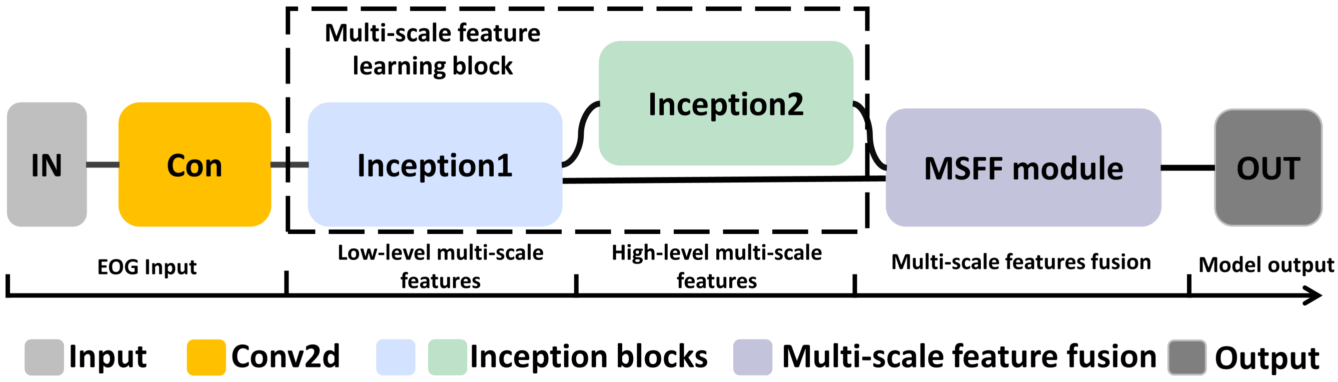
Multi-scale feature learning block
IN
Con
Inception1
Inception2
MSFF module
OUT
EOG Input
Low-level multi-scale features
High-level multi-scale features
Multi-scale features fusion
Model output
Input
Conv2d
Inception blocks
Multi-scale feature fusion
Output

Fig. 1. The overall framework of the MIDF-NET.

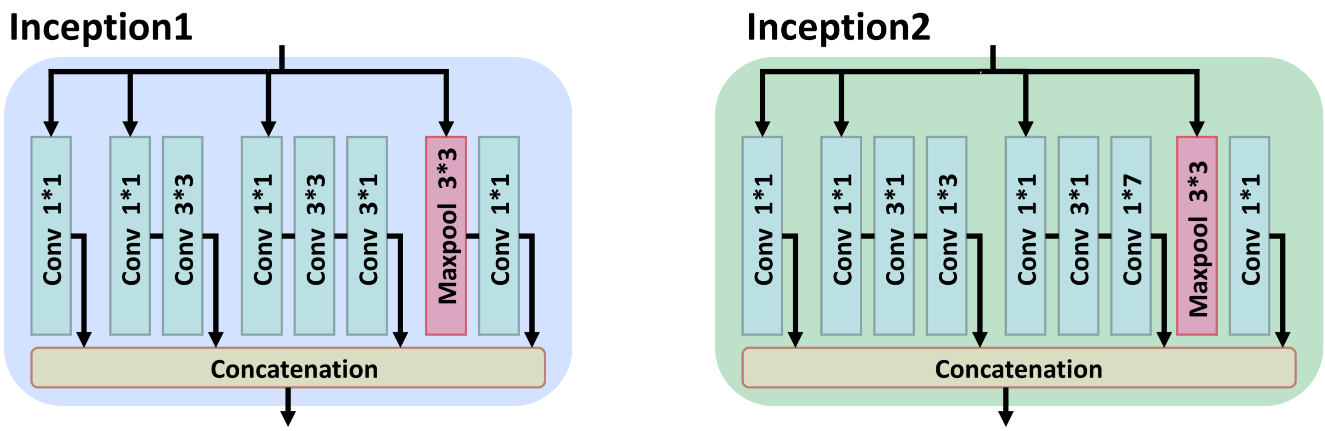

Fig. 2. The structure of the Inception blocks. The Inception Block 1 extracts the low-level multi-scale features; the Inception Block 2 extracts the high-level multi-scale features.

There is still room for DL architecture improvement. To address these flaws, this paper presents a novel MIDF-NET framework to improve the performance of EOG-based EMC. We pursue the idea that the enhancement of deep features can significantly increase the representational ability of the DL model. The MIDF-NET can synchronously benefit from low-level and high-level multi-scale features to well distinguish diverse eye movements, resulting in superior performance for EOG-based EMC.

## III. Materials and Methods

### A. Datasets and data pre-processing

As shown in Table II, the EOG data was collected from five studies with a total of 50 subjects and 59 recordings [18]–[22]. These studies have their own tasks, and the EOG signals are acquired from the calibration phase during which the subject moves their eyes. According to the EOG signals, experimenters labeled each data point as eye blink, rightward eye movement, leftward eye movement, upward eye movement, downward eye movement, or resting. Since the subjects may exhibit certain body movements during the rest period, and because body movements are not eye movements but can significantly interfere with EOG information, these movements are considered noise for EOG signals. Therefore, we mainly extract five types of eye movements (excluding resting) to evaluate model performance.

According to the literature, experimenters from all five databases conducted screenings of the raw data to ensure the quality of data within the databases. A Butterworth notch filter with 49Hz and 51Hz cut-off frequencies and a high-pass filter with 0.4Hz cut-off frequency are applied to attenuate line noise and drifts. EOG electrodes placed next to the outer canthi of both eyes are used to compute the horizontal EOG (HEOG) derivative. The vertical EOG (VEOG) derivative is computed using the electrode next to the superior and inferior orbital [23]. Regarding the EOG dataset 3, the VEOG derivative was set to the mean of the differences between the electrode above the nasion and the ones below the outer canthi as described in [24]. The EOG signals are subsequently low-pass filtered with a cut-off 5Hz cut-off frequency. Since the five datasets employ different sampling frequencies, we first downsample the EOG signals to 100 Hz mutually. The EOG signal segments are extracted according to the data point label of different eye movements. To be specific, one eye movement sample (right, left, up, down, and blink) is represented using 1s HEOG and VEOG segments. If a sample is insufficient or exceeds the data length of one second, this sample will be completed using the last data point or truncated directly. In the end, we combine the preprocessed HEOG and VEOG data under the corresponding eye movement labels into a 2x100 tensor. This tensor serves as the input to the designed network, with HEOG and VEOG data forming the first and second rows, respectively.

The five public EOG datasets: https://osf.io/2qgrd/wiki/home/

TABLE I
RELATED WORKS

| Study | Eye movement approach | Models | Evaluation approach | Number of eye movement | Subjects number |
|---|---|---|---|---|---|
| Lopez *et al.* [8] | ML-based | KNN, SVM, ANN, Voting, and Two-stage ensemble model | Cross-subject | 5 | 15 |
| O'Board *et al.* [9] | ML-based | DT, SVM, and KNN | Intra-subject | 5 | 1 |
| Bulling *et al.* [10] | ML-based | SVM | Cross-subject | 3 | 8 |
| Zou *et al.* [15] | DL-based | Single-scale CNN | Intra-subject | 150 words* | 6 |
| Aya *et al.* [16] | DL-based | LSTM | Intra-subject | 4 | 6 |
| This study | DL-based | Multi-scale CNN | Cross-subject | 5 | 50 |

150 words*: In [15], the classification task is to classify 150 eye movement sequences that represent 150 words.

TABLE II
THE 5 DATASETS USED IN THIS STUDY

| Dataset | Number of subjects | Number of EOG elecrodes | Sampling rates (Hz) | Number of samples | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | | **Blink** | **Up** | **Down** | **Right** | **Left** |
| EOG dataset 1 [18] | 5 | 6 | 200 | 331 | 267 | 259 | 238 | 229 |
| EOG dataset 2 [19] | 15 | 6 | 200 | 1018 | 649 | 664 | 670 | 663 |
| EOG dataset 3 [20] | 10 | 6 | 200 | 905 | 431 | 390 | 396 | 416 |
| EOG dataset 4 [21] | 15 | 3 | 100 | 1373 | 648 | 661 | 537 | 489 |
| EOG dataset 5 [22] | 14 | 6 | 256 | 1276 | 759 | 793 | 986 | 916 |

## B. Proposed networks

The overall framework of the proposed network is shown in Fig. 1. Our proposed framework mainly consists of three parts: (1) a multi-scale feature learning block that utilizes two Inception modules with different temporal scales to extract multi-scale features from EOG signals. (2) A feature fusion block that fuses features at different layers, taking advantage of the high and low-level multi-scale information. (3) a shallow network that contains 2 fully connected layers to decode the specific eye movement. Here, we give a brief description of these blocks.

*Multi-scale feature learning block*: This block contains two Inception modules (denoted as Incp1 and Incp2) for multi-scale feature learning (As shown in Fig. 2). Each Inception block consists of four CNN streams with different kernel scales. The ReLU function is applied to introduce the nonlinearity.

In the design of the Incp1, a different-depth CNN architecture was adopted to extract deep features at different scales. Given that prior research has demonstrated the effectiveness of a CNN architecture with concatenated small receptive fields in feature extraction compared to a single large receptive field, we incorporated this idea to enhance the MIDF-NET [25]. In addition, Incep1 also adopted the concept of 'Network in Network' to enable information fusion between channels and reduce computational complexity [26].

In the design of Incp2, we followed previous studies that reported advantages in performance improvement and parameter reduction by placing an asymmetric architecture in the middle of networks [26]. Hence, we adopted an asymmetric convolutional Inception module to further extract the multi-scale representation at a high level.

The multi-scale features generated by different streams in each Inception block are then concatenated and serve as the input for the next layer.

*Multi-scale Feature Fusion (MSFF) module*: Apart from the Multi-scale feature learning, we utilized a novel MSFF module to fuse the learned multi-scale features. Specifically, we employed this feature fusion block to further fuse the multi-scale features generated by Incp1 and Incp2 modules. This fusion strategy was inspired by [27], since in this way the proposed network may combine low and high-level multi-scale features to infer the eye movement patterns, thus taking full advantage of detailed and semantic information. The specific structure of the MSFF module is shown in Fig. 3. To avoid the magnitude variations between multi-scale features extracted from different channels, a batch normalization (BN) layer was applied for normalization. Suppose the multi-scale features extracted by the $i_{th}$ Inception module is $M_i$, the fused features can be represented by the following formula:

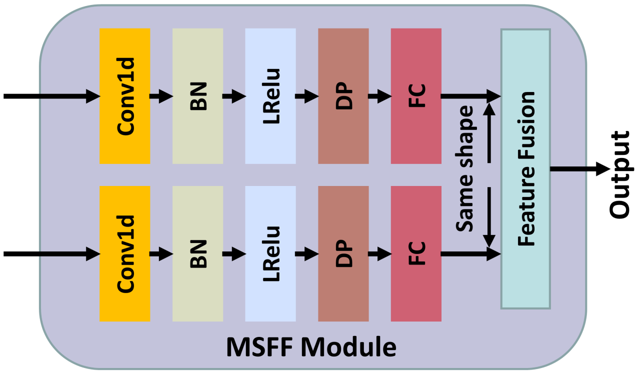

Fig. 3. The structure of the MSFF module.

$$M_{fused} : H_{low}(M_1) + H_{high}(M_2) \tag{1}$$

where the $H_{low}$ and $H_{high}$ correspond to the low and high-level feature fusion, respectively.

*Shallow network:* Here, two fully connected (FC) layers were used to reduce the dimension of the vector gradually. The softmax layer was then computed as the predicted probability of the specific eye movement. Further, The cross-entropy loss was employed as the network loss to measure the consistency between the predicted label distribution and the real label distribution.

## C. Hyper-parameters setting and evaluation metrics

We evaluate the performance using the Leave-one-subject cross-validation strategy. Specifically, in each dataset, recordings from $N_s-1$ subjects are used to train the model, and the recordings from the remaining one's data serve as the test set. Here, $N_s$ is equal to the subject number of each dataset. This process is repeated $N_s$ times to ensure each subject's data serves as the test set once. Furthermore, we extracted 20% of the data of each subject in the training set to

TABLE III

SUMMARY OF THE HYPER-PARAMETERS

| Parameters | Values |
|---|---|
| Batch size | 256 |
| Maximum epochs | 200 |
| learning rate | 0.008 |
| Dropout probability | 0.2 |
| StepLR $\gamma$ | 0.3 |
| StepLR step size | 40 |
| Weight decay | 0.0004 |

construct the validation set (training data: validation data = 8:2). The trained model with the highest accuracy on the validation set is used for testing.

The specific hyper-parameters of the proposed network are shown in Table III. To illustrate, we used the Adam optimizer to update the network [28]. The initial learning rate was set as $8 * 10^{-3}$ empirically. For the learning rate adjustment strategy, we applied the stepLR techniques to attenuate the learning rate to promote the network to converge to the optimal solution better. All the dropout layers were defined with a probability $p = 0.2$ to prevent overfitting problems [29]. We set the maximum number of epochs to 200 and the optimal trained model was selected according to the performance on the validation set for further testing.

Since the quantity of various eye movement data is relatively balanced, the network performance is mainly evaluated via the accuracy (Acc) metrics.

$$Acc = \frac{1}{N}\sum_{i=0}^{N} TP_i \tag{2}$$

In addition, specificity (Spe), sensitivity (Sen), and per-class accuracy (P-class acc) are also calculated for auxiliary evaluation.

$$Spe = \frac{1}{N}\sum_{i=0}^{N} \frac{TN_i}{TN_i + FP_i} \tag{3}$$

$$Sen = \frac{1}{N}\sum_{i=0}^{N} \frac{TP_i}{TP_i + FN_i} \tag{4}$$

$$P - class \quad acc = \frac{TP_i}{N_i} \tag{5}$$

where $N$ and $N_i$ are the total numbers of samples and $i_{th}$ class samples. True Positives ($TP_i$), False Positives ($FP_i$), True Negatives ($TN_i$), and False Negatives ($FN_i$) represent the predicted number of positive and negative samples that are correct or incorrect for the i-class eye movements.

### D. Methods for comparison

Prevailing EOG-based approaches were implemented on the five datasets for comparison. Here, we give a brief illustration of the comparative approaches.

• ML-based approaches: For the ML methods, we selected SVM, Two-stage ensemble, KNN, decision trees, and voting models which have shown superior performance in [8], [9]. The corresponding hand-crafted features and parameter settings are consistent with related works.

• DL-based approaches: Regarding the DL methods, a recently-published single-scale CNN (SS-CNN), named the eye-say network, was implemented [15]. Besides, referring to the concept of parallel multi-scale CNN (PMS-CNN) [30]–[33], we also implemented a PMS-CNN network for comparison. PMS-CNN can extract the multi-scale features from EOG signals in a parallel layer-wise manner. We conducted comprehensive experiments to determine the structure of PMS-CNN by exploring kernel sizes, number of branches, and number of layers. The final structure of PMS-CNN has three branches with kernel size $[1, 3, 5]$, each branch with three layers of CNNs.

### E. Statistical analysis

Statistical analysis was performed on the accuracy metric. The Shapiro-Wilk test was first performed to verify the normality of data. Then, the homogeneity of variance was tested. If the results meet the above requirements, a one-way repeated-measures (RM) analysis of variance (ANOVA) was applied. Otherwise, A non-parametric test was applied for statistical analysis. For the non-parametric test, the Friedman test was performed to detect whether there was an overall significant difference. If the overall statistical difference was significant, the Wilcoxon signed-rank test was further conducted to perform the pair-wise comparison. To avoid errors caused by multiple comparisons, the Holm-Bonferroni correction was adopted. The differences were considered significant if $p < 0.05$ was achieved.

## IV. RESULT

### A. Performance of the MIDF-NET

Table IV shows the performance of the MIDF-NET across Acc, Spe, Sen, and P-class acc metrics. The bold values in each column denote the highest performance. Not surprisingly, the MIDF-NET achieves the highest performance across the 5 datasets (Acc: 87.7%, 86.0%, 95.0%, 94.2%, and 95.4%). In addition, the MIDF-NET maintains stable accuracy in recognizing multiple eye movements, indicating its robustness without bias for a single category. Overall, MIDF-NET shows better EMC performance on multiple datasets. Under different conditions (different subjects, electrode schemes, and data acquisition scenes), MIDF-NET can effectively decode accurate eye movement patterns from raw EOG signals.

### B. Ablation Experiments

To better understand the MIDF-NET, we performed ablation experiments to fully analyze the architecture of MIDF-NET. Concretely, we derived three model variants: 1) Incep1 block only; 2) Incep1 + Incep2 blocks; 3) Incep1 + Incep2 + Multi-scale feature fusion module (The proposed model).

As shown in Fig. 4, merely using low-level (Incep1 block only) or high-level (Incep1 + Incep2 blocks) multi-scale features results in similar classification performance. By contrast, as we further fuse the low-level and high-level multi-scale features (Incep1 + Incep2 + Multi-scale feature fusion module), we observe a significant improvement in classification accuracy, achieving accuracies of 87.7%, 86.0%, 95.0%, 94.2%, and 95.4% on five databases, respectively. This result confirms our hypothesis. By fusing the multi-scale features properly, the MIDF-NET can take advantage of the complementary information from low and high-level features, thereby inferring eye movements effectively.

### C. Comparison with prevailing ML/DL methods

Table IV also presents the comparison between MIDF-NET and state-of-the-art ML methods. Compared to the best ML models on the 5 datasets, the improved accuracy can be up to 2.7%, 7.6%, 3.1%, 3.4%, and 2.0%, respectively. The P-class accuracy also demonstrates that MIDF-NET achieves average and high accuracy in classifying separate categories, whereas the ML methods show a certain degree of model bias for the rightward eye movement. Statistical analysis

TABLE IV
PERFORMANCE COMPARISON WITH STATE-OF-THE-ART ML METHODS

| Dataset | Models | Acc | Sen | Spe | P-class acc | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | Right | Left | Up | Down | Blink |
| EOG dataset1 | Two Stage Ensemble [8] | 84.4% | 83.2% | 96.1% | **99.0%** | 85.6% | 75.8% | 77.4% | 78.2% |
| | Voting [8] | 85.0% | 84.0% | 96.3% | 97.2% | 87.9% | 77.6% | 75.5% | 81.9% |
| | SVM [9] | 79.9% | 78.8% | 95.0% | 95.5% | 73.7% | 66.2% | 80.0% | 78.7% |
| | KNN [9] | 84.1% | 82.9% | 96.0% | 97.7% | **91.6%** | 74.4% | 73.0% | 77.9% |
| | Decision Tree [9] | 81.7% | 80.4% | 95.5% | 98.2% | 78.4% | 74.8% | 68.1% | 82.4% |
| | **MIDF-NET** | **87.7%** | **87.4%** | **96.9%** | 86.9% | 78.8% | **89.4%** | **89.7%** | **92.1%** |
| EOG dataset2 | Two Stage Ensemble [8] | 77.4% | 74.9% | 94.4% | 99.3% | 75.9% | 70.1% | 65.9% | 63.2% |
| | Voting [8] | 78.4% | 76.3% | 94.7% | 98.1% | 74.2% | 72.3% | 69.2% | 67.7% |
| | SVM [9] | 75.3% | 72.8% | 93.9% | 98.3% | 62.9% | 66.9% | 68.5% | 67.2% |
| | KNN [9] | 77.5% | 75.0% | 94.5% | **99.4%** | 76.6% | 66.1% | 66.0% | 66.8% |
| | Decision Tree [9] | 73.4% | 70.9% | 93.5% | 96.6% | 63.3% | 63.9% | 65.2% | 65.7% |
| | **MIDF-NET** | **86.0%** | **84.8%** | **96.5%** | 79.2% | **78.1%** | **84.0%** | **83.3%** | **99.3%** |
| EOG dataset3 | Two Stage Ensemble [8] | 90.2% | 88.2% | 97.6% | **99.1%** | 86.3% | 89.1% | 87.7% | 78.6% |
| | Voting [8] | 91.9% | 90.3% | 98.0% | 98.5% | 85.2% | **92.0%** | 89.7% | 86.1% |
| | SVM [9] | 89.0% | 87.6% | 97.3% | 95.3% | 80.3% | 90.7% | 87.9% | 83.5% |
| | KNN [9] | 90.8% | 89.0% | 97.8% | 98.9% | 89.6% | 87.3% | 88.5% | 81.0% |
| | Decision Tree [9] | 87.4% | 85.6% | 96.9% | 95.8% | 76.8% | 89.6% | 88.6% | 77.0% |
| | **MIDF-NET** | **95.0%** | **94.2%** | **98.8%** | 94.3% | **96.3%** | 91.2% | **90.6%** | **98.6%** |
| EOG dataset4 | Two Stage Ensemble [8] | 89.8% | 88.6% | 97.4% | 92.1% | 90.3% | 89.4% | 87.9% | 83.2% |
| | Voting [8] | 90.4% | 89.1% | 97.5% | 92.0% | 91.7% | 88.7% | 87.4% | 85.9% |
| | SVM [9] | 86.7% | 84.1% | 96.7% | 93.8% | 77.8% | 87.8% | 83.3% | 77.8% |
| | KNN [9] | 90.6% | 89.2% | 97.6% | 92.7% | 92.1% | 88.0% | 87.6% | 85.5% |
| | Decision Tree [9] | 90.8% | 89.4% | 97.7% | **94.4%** | 88.5% | 90.2% | 87.4% | 86.6% |
| | **MIDF-NET** | **94.2%** | **93.6%** | **98.5%** | 92.7% | **93.3%** | **93.3%** | **95.6%** | **93.0%** |
| EOG dataset5 | Two Stage Ensemble [8] | 91.6% | 90.4% | 97.9% | 95.4% | 91.2% | 95.2% | 91.8% | 78.4% |
| | Voting [8] | 93.4% | 92.4% | 98.4% | 95.9% | 93.0% | 95.2% | 92.1% | 86.0% |
| | SVM [9] | 91.7% | 90.4% | 97.9% | **96.2%** | 87.0% | **95.6%** | 91.9% | 81.5% |
| | KNN [9] | 92.1% | 90.9% | 98.0% | 95.4% | **94.8%** | 95.0% | 92.8% | 76.5% |
| | Decision Tree [9] | 90.8% | 89.3% | 97.7% | 94.3% | 86.6% | 94.0% | 91.4% | 80.4% |
| | **MIDF-NET** | **95.4%** | **94.9%** | **98.8%** | 94.3% | 91.8% | 91.8% | **93.0%** | **98.1%** |

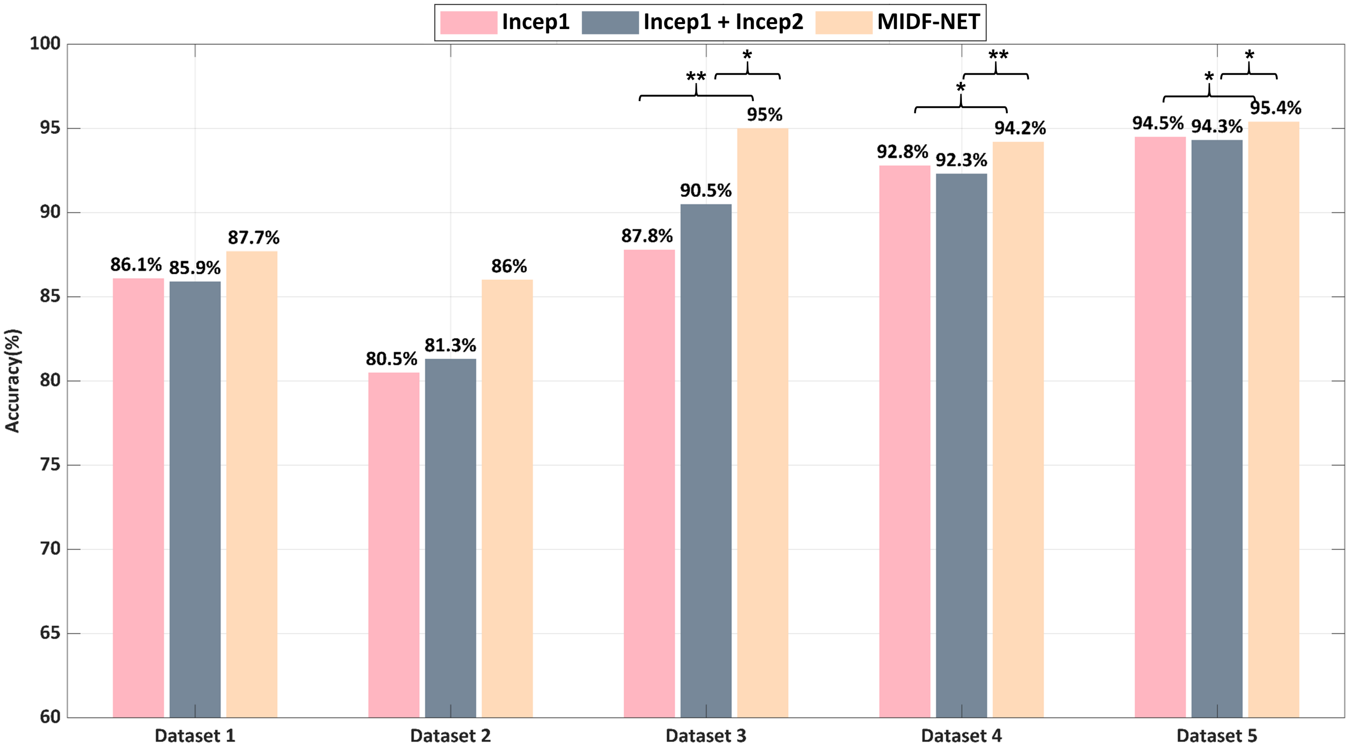

Fig. 4. The ablation results of different network architectures. *:$p$<0.05,**:$p$<0.01

demonstrates that the MIDF-NET significantly outperforms all ML-based methods on EOG dataset2 ($p$ <0.01), EOG dataset3 ($p$ <0.05), EOG dataset4 ($p$ <0.01), and EOG dataset5 ($p$ <0.01).

As shown in Fig. 5, the SS-CNN achieves an accuracy of 73.9%, 74.2%, 82.4%, 87.1%, and 91.3% on the 5 EOG datasets. By contrast, the MIDF-NET outperforms the SS-CNN with an accuracy improvement of 13.9%, 11.8%, 12.6%, 7.1%, and 4.1%. Statistical analysis demonstrates that MIDF-NET significantly outperforms SS-CNN on EOG Dataset2 ($p$=0.008), EOG Dataset3 ($p$=0.04), EOG Dataset4 ($p$=0.0054), and EOG Dataset5 ($p$=0.0045).

Compared with PMS-CNN, our proposed MIDF-NET still outperforms PMS-CNN with an accuracy improvement of 1.9%, 6.6%, 6.4%, 3.3%, and 0.6%. Statistic analysis showes that MIDF-NET surpassed PMS-CNN significantly on EOG Dataset2 ($p$=0.007), EOG Dataset3 ($p$=0.035), EOG Dataset4 ($p$=0.0054), and EOG Dataset5 ($p$=0.0045).

## V. Discussion

### A. Comparison with existing works

The DL methods can automatically learn the discriminative features from raw signals. In contrast, the apriori knowledge provided to the ML models needs to be carefully selected, as these hand-crafted features often play a crucial role in ML performance. The MIDF-NET outperforms all ML approaches across the 5 datasets. One possible explanation may be that the MIDF-NET method can take advantage of a large amount of data to learn the representative representations. Since the 5 EOG datasets used in this study cover different electrode placements, acquisition situations, and subject variants, a single hand-crafted feature set may be insufficient to well present the multiple eye movements. Besides, although diverse feature optimization techniques can be utilized to generate the enhanced feature subset, the robustness is still lacking when facing complex situations with multiple datasets. In comparison, the MIDF-NET can automatically adapt the model parameters from sufficient training data. Guided by well-designed learning tasks, the MIDF-NET can effectively extract discriminative multi-scale features from multiple EOG datasets.

Involving multiple scales in CNNs has shown great advantages in time-series signals processing [30], [34]. This idea first appeared in Googlenet and was denoted as ”Inception blocks” [35]. Since multi-scale analyses have been proven to the strong candidates for successful EOG-based EMC [13], [36], we also adopted this effective concept to ensure a powerful feature learning capacity for MIDF-NET. Compared with the traditional SS-CNN or PMCS CNN architecture, the inception blocks can extract more discriminative multi-scale representations. At the same time, the design of decomposing convolutional blocks into asymmetric convolutions can also effectively reduce network parameters to control the risk of overfitting. However, to our best knowledge, very few studies have utilized this module for EOG-based EMC. Based on the results, the effectiveness of this module was first confirmed. We hope it can inspire diverse multi-scale solutions in the future.

In addition, fusing features from different levels has been an important strategy in many deep learning tasks [37]–[39]. In our work, the MIDF-NET also benefits from this concept to utilize complementary yet correlated information from low and high-level multi-scale features. Commonly, the low-level features are rich in detailed information, whereas the high-level features are rich in semantic information. To generate semantic and detailed representations, the MIDF-NET integrates the multi-scale feature fusion module to take advantage of the complementary information from low and high-level features. Such operations bring two obvious advantages [40]. First, the fused features may reduce the detailed information loss as the EOG signals flow through the networks. As the low-level features (e.g., shape and details) often serve as the basis for classification in EOG-based eye movement tasks, preserving detailed information is advantageous for the final EMC. Second, a direct combination of low and high-level features can only bring limited improvement without considering feature fusion [41]. Due to the gap in high and low levels, the simple combination is often insufficient and may involve excessive irrelevant features. On the contrary, the multi-scale feature fusion module in MIDF-NET can further refine the features, which avoids involving redundant features and enables the model to better integrate low and high-level features. As the results in Fig. 4, we can see that the utilization of MSFF modules brings significant improvements. The MIDF-NET can actually, to an extent, automatically supplement the missing detail information of high-level semantic features to facilitate the model capacity in classifying multiple eye movements.

### B. MIDF-NET for EOG-HCI

The purpose of this study is to develop a classifier for EOG-HCI research that is generalizable and robust to meet various interaction design requirements. Existing EOG-HCI research heavily relies on the specific data acquired, and the structure of classification models often needs to be tailored to classification tasks. This often results in low generalization and robustness of the models.

Therefore, using the models proposed in existing EOG-HCI research has adverse effects on the development of user experience and interaction design work for the EOG-HCI system. From the perspective of user experience, previous EMC models have poor adaptability to new data, especially data collected in non-laboratory environments. The performance of eye movement recognition is often unstable. To ensure accuracy, developers have to make trade-offs between accuracy and interaction efficiency (such as pre-calibration for each individual or requiring users to confirm multiple times to ensure the accuracy of interaction commands), ultimately resulting in a poor user experience. From the perspective of interaction design work, since the performance of previous EMC models are closely related to the method of EOG signal collection, developers may find it difficult to adjust electrode configurations based on design requirements when using these models for interaction design. This is the case even if certain electrode configurations affect the user experience. Relying on the network structure’s ability to capture the inherent characteristics of eye movements, MIDF-NET can perform well under different data conditions, providing a stable algorithmic foundation for various EOG-HCI studies. This enables:

- MIDF-NET has a relatively accurate recognition capability across different users, ensuring that HCI systems based on the MIDF-NET algorithm can accurately and stably recognize user interaction commands through eye movements without relying on pre-calibration.
- During the training process of the network, data from various electrode configurations were used. This makes the recognition performance of MIDF-NET independent of specific electrode setups. Therefore, when developing HCI systems using the MIDF-NET algorithm, developers do not need to be overly concerned with the electrode configuration of the EOG signals. This allows designers to focus more on the user experience itself, especially when designing wearable EOG-HCI interaction systems.

### C. Limitations and future work

Although the MIDF-NET shows promising performance in classifying multiple eye movements, this work can still be enhanced via the following improvements:

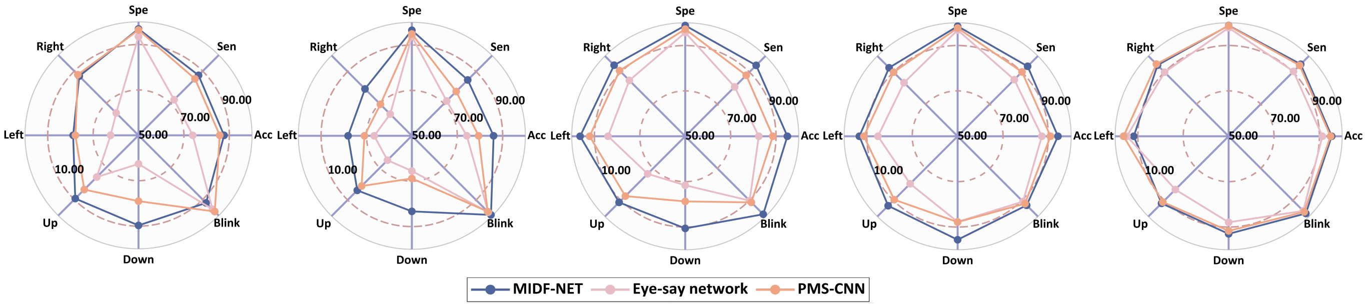

Fig. 5. The comparison result between the SS-CNN, PMS-CNN, and our proposed MIDF-NET. From left to right the radar figures correspond to EOG dataset1, EOG dataset2, EOG dataset3, EOG dataset4, and EOG dataset5.

1) Designing novel feature fusion approaches: We assume that fusing low- and high-level features is beneficial for the model to distinguish different eye movement patterns. In the future, more feature fusion approaches including attention mechanisms can be investigated to well capture the channel or spatial information.

2) Classifying more eye movements: In our study, we mainly classify 5 frequently used eye movement patterns (right, left, up, down, and blink). However, there are still many complex patterns of eye movement that are worth exploring. In the future, we aim to build a multiple-eye movement dataset and evaluate the MIDF-NET on a more challenging task.

3) Real-time-oriented algorithms: In HCI scenarios, interaction response time is a crucial metric for user experience. However, this study did not investigate the performance of MIDF-NET in low response time real-time eye movement classification (EMC) scenarios. Therefore, in our future research, we will focus on enhancing MIDF-NET for EMC tasks in online real-time usage scenarios to achieve objective EMC performance within shorter response times.

4) Improving the model's noise and artifact robustness: EOG signals are susceptible to interference from noise or artifacts, leading to the loss of effective information. involuntary blinking and subtle facial expressions during eye movements can affect the performance of current algorithms, ultimately resulting in the misidentification of user actions as incorrect interaction commands, thereby impacting the usability and user experience of the entire interaction system. Therefore, in future work, we will focus on enhancing the algorithm's noise robustness.

## VI. Conclusion

In this study, we proposed a novel convolution neural network, called MIDF-NET, for accurate EOG-based EMC. The MIDF-NET efficiently integrates two inception modules and an MSFF module to distinguish the 5 basic eye movements (right, left, up, down, and blink). The Inception modules facilitate the extraction of multi-scale features at various temporal scales. The fusion subnetwork fully fuses the feature to take advantage of multi-scale representations at different layers. We comprehensively evaluate the MIDF-NET on 5 public datasets which involve 50 subjects and 59 recordings. The proposed MIDF-NET can achieve excellent performance, significantly outperforming state-of-the-art methods with an average accuracy of 92.16% across 5 datasets. In addition, we also conducted ablation experiments to better analyze the proposed model. Since the MIDF-NET can effectively extract and fuse the multi-scale features from EOGs, it is expected to broaden the horizons of future study of EOG-based eye movement model design.

## Acknowledgement

The authors would like to thank Professor Loe Feijs for providing valuable suggestions on human-computer interaction and comments on earlier drafts of this paper.

## References

[1] H. S. Dhillon, R. Singla, N. S. Rekhi, and R. Jha, "Eog and emg based virtual keyboard: A brain-computer interface," in *2009 2nd IEEE International Conference on Computer Science and Information Technology*. IEEE, 2009, pp. 259–262.

[2] A. B. Usakli and S. Gurkan, "Design of a novel efficient human–computer interface: An electrooculagram based virtual keyboard," *IEEE transactions on instrumentation and measurement*, vol. 59, no. 8, pp. 2099–2108, 2009.

[3] N. Barbara, T. A. Camilleri, and K. P. Camilleri, "Eog-based eye movement detection and gaze estimation for an asynchronous virtual keyboard," *Biomedical Signal Processing and Control*, vol. 47, pp. 159–167, 2019.

[4] P. Zhang, M. Ito, S.-i. Ito, and M. Fukumi, "Implementation of eog mouse using learning vector quantization and eog-feature based methods," in *2013 IEEE Conference on Systems, Process & Control (ICSPC)*. IEEE, 2013, pp. 88–92.

[5] Z. Zeng, L. Tao, R. Su, Y. Zhu, L. Meng, A. Tuheti, H. Huang, F. Shu, W. Chen, and C. Chen, "Unsupervised transfer learning approach with adaptive reweighting and resampling strategy for inter-subject eog-based gaze angle estimation," *IEEE Journal of Biomedical and Health Informatics*, 2023.

[6] K.-R. Lee, W.-D. Chang, S. Kim, and C.-H. Im, "Real-time "eye-writing" recognition using electrooculogram," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 25, no. 1, pp. 37–48, 2016.

[7] F. D. Perez Reynoso, P. A. Niño Suarez, O. F. Aviles Sanchez, M. B. Calva Yañez, E. Vega Alvarado, and E. A. Portilla Flores, "A custom eog-based hmi using neural network modeling to real-time for the trajectory tracking of a manipulator robot," *Frontiers in Neurorobotics*, vol. 14, p. 578834, 2020.

[8] A. López, J. R. Villar, M. Fernández, and F. J. Ferrero, "Comparison of classification techniques for the control of eog-based hcis," *Biomedical Signal Processing and Control*, vol. 80, p. 104263, 2023.

[9] B. O'Bard and K. George, "Classification of eye gestures using machine learning for use in embedded switch controller," in *2018 IEEE International Instrumentation and Measurement Technology Conference (I2MTC)*. IEEE, 2018, pp. 1–6.

[10] A. Bulling, J. A. Ward, H. Gellersen, and G. Tröster, "Eye movement analysis for activity recognition using electrooculography," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 4, pp. 741–753, 2010.

[11] S. Mala and K. Latha, "Feature selection in classification of eye movements using electrooculography for activity recognition," *Computational and mathematical methods in medicine*, vol. 2014, 2014.

[12] L. J. Qi and N. Alias, "Comparison of ann and svm for classification of eye movements in eog signals," in *Journal of Physics: Conference Series*, vol. 971, no. 1. IOP Publishing, 2018, p. 012012.

[13] S. I. Khan and R. B. Pachori, "Automated eye movement classification based on emg of eom signals using fbse-ewt technique," *IEEE Transactions on Human-Machine Systems*, vol. 53, no. 2, pp. 346–356, 2023.

[14] Z. Lv, Y. Wang, C. Zhang, X. Gao, and X. Wu, "An ica-based spatial filtering approach to saccadic eog signal recognition," *Biomedical Signal Processing and Control*, vol. 43, pp. 9–17, 2018.

[15] J. Zou and Q. Zhang, "eyesay: Brain visual dynamics decoding with deep learning & edge computing," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 30, pp. 2217–2224, 2022.

[16] A. R. Abih and M. J. Hayawi, "Comparison of lstm and svm for classification of eye movements in eog signals," *Journal of Al-Qadisiyah for computer science and mathematics*, vol. 14, no. 3, pp. Page–130, 2022.

[17] G. Jialu, S. Ramkumar, G. Emayavaramban, M. Thilagaraj, V. Muneeswaran, M. P. Rajasekaran, and A. F. Hussein, "Offline analysis for designing electrooculogram based human computer interface control for paralyzed patients," *IEEE Access*, vol. 6, pp. 79 151–79 161, 2018.

[18] R. J. Kobler, A. I. Sburlea, and G. R. Müller-Putz, "A comparison of ocular artifact removal methods for block design based electroencephalography experiments." in *GBCIC*, 2017.

[19] ——, "Tuning characteristics of low-frequency eeg to positions and velocities in visuomotor and oculomotor tracking tasks," *Scientific reports*, vol. 8, no. 1, pp. 1–14, 2018.

[20] V. Mondini, R. J. Kobler, A. I. Sburlea, and G. R. Müller-Putz, "Continuous low-frequency eeg decoding of arm movement for closed-loop, natural control of a robotic arm," *Journal of Neural Engineering*, vol. 17, no. 4, p. 046031, 2020.

[21] C. Lopes-Dias, A. I. Sburlea, and G. R. Müller-Putz, "Online asynchronous decoding of error-related potentials during the continuous control of a robot," *Scientific reports*, vol. 9, no. 1, pp. 1–9, 2019.

[22] A. Schwarz, J. Pereira, R. Kobler, and G. R. Müller-Putz, "Unimanual and bimanual reach-and-grasp actions can be decoded from human eeg," *IEEE transactions on biomedical engineering*, vol. 67, no. 6, pp. 1684–1695, 2019.

[23] R. J. Kobler, A. I. Sburlea, C. Lopes-Dias, A. Schwarz, M. Hirata, and G. R. Müller-Putz, "Corneo-retinal-dipole and eyelid-related eye artifacts can be corrected offline and online in electroencephalographic and magnetoencephalographic signals," *NeuroImage*, vol. 218, p. 117000, 2020.

[24] A. Schlögl, C. Keinrath, D. Zimmermann, R. Scherer, R. Leeb, and G. Pfurtscheller, "A fully automated correction method of eog artifacts in eeg recordings," *Clinical neurophysiology*, vol. 118, no. 1, pp. 98–104, 2007.

[25] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.

[26] M. Lin, Q. Chen, and S. Yan, "Network in network," *arXiv preprint arXiv:1312.4400*, 2013.

[27] W. Wei, Q. Dai, Y. Wong, Y. Hu, M. Kankanhalli, and W. Geng, "Surface-electromyography-based gesture recognition by multi-view deep learning," *IEEE Transactions on Biomedical Engineering*, vol. 66, no. 10, pp. 2964–2973, 2019.

[28] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[29] S. Jégou, M. Drozdzal, D. Vazquez, A. Romero, and Y. Bengio, "The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2017, pp. 11–19.

[30] G. Hajian and E. Morin, "Deep multi-scale fusion of convolutional neural networks for emg-based movement estimation," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 30, pp. 486–495, 2022.

[31] L. Shen, M. Sun, Q. Li, B. Li, Z. Pan, and J. Lei, "Multiscale temporal self-attention and dynamical graph convolution hybrid network for eeg-based stereogram recognition," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 30, pp. 1191–1202, 2022.

[32] P. Thuwajit, P. Rangpong, P. Sawangjai, P. Autthasan, R. Chaisaen, N. Banluesombatkul, P. Boonchit, N. Tatsaringkansakul, T. Sudhawiyangkul, and T. Wilaiprasitporn, "Eegwavenet: Multiscale cnn-based spatiotemporal feature extraction for eeg seizure detection," *IEEE Transactions on Industrial Informatics*, vol. 18, no. 8, pp. 5547–5557, 2021.

[33] B. Sun, B. Song, J. Lv, P. Chen, X. Sun, C. Ma, and Z. Gao, "A multi-scale feature extraction network based on channel-spatial attention for electromyographic signal classification," *IEEE Transactions on Cognitive and Developmental Systems*, 2022.

[34] A. H. Ansari, K. Pillay, A. Dereymaeker, K. Jansen, S. Van Huffel, G. Naulaers, and M. De Vos, "A deep shared multi-scale inception network enables accurate neonatal quiet sleep detection with limited eeg channels," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 3, pp. 1023–1033, 2021.

[35] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[36] R. Barea, L. Boquete, S. Ortega, E. López, and J. Rodríguez-Ascariz, "Eog-based eye movements codification for human computer interaction," *Expert Systems with Applications*, vol. 39, no. 3, pp. 2677–2683, 2012.

[37] K. Li, C. Zou, S. Bu, Y. Liang, J. Zhang, and M. Gong, "Multi-modal feature fusion for geographic image annotation," *Pattern recognition*, vol. 73, pp. 1–14, 2018.

[38] X. Jin, Q. Xiong, C. Xiong, Z. Li, and Z. Gao, "Single image super-resolution with multi-level feature fusion recursive network," *Neurocomputing*, vol. 370, pp. 166–173, 2019.

[39] Y. Fang, S. Gao, J. Li, W. Luo, L. He, and B. Hu, "Multi-level feature fusion based locality-constrained spatial transformer network for video crowd counting," *Neurocomputing*, vol. 392, pp. 98–107, 2020.

[40] X. Li, D. Song, and Y. Dong, "Hierarchical feature fusion network for salient object detection," *IEEE Transactions on Image Processing*, vol. 29, pp. 9165–9175, 2020.

[41] E. Elizar, M. A. Zulkifley, R. Muharar, M. H. M. Zaman, and S. M. Mustaza, "A review on multiscale-deep-learning applications," *Sensors*, vol. 22, no. 19, p. 7384, 2022.