Computers, Materials & Continua

Tech Science Press

DOI: 10.32604/cmc.2024.055768

ARTICLE

IDSSCNN-XgBoost: Improved Dual-Stream Shallow Convolutional Neural Network Based on Extreme Gradient Boosting Algorithm for Micro Expression Recognition

Adnan Ahmad, Zhao Li*, Irfan Tariq and Zhengran He

School of Information Science and Engineering, Southeast University, Nanjing, 210016, China

*Corresponding Author: Zhao Li. Email: zhaoli@seu.edu.cn

Received: 06 July 2024 Accepted: 18 October 2024 Published: 03 January 2025

ABSTRACT

Micro-expressions (ME) recognition is a complex task that requires advanced techniques to extract informative features from facial expressions. Numerous deep neural networks (DNNs) with convolutional structures have been proposed. However, unlike DNNs, shallow convolutional neural networks often outperform deeper models in mitigating overfitting, particularly with small datasets. Still, many of these methods rely on a single feature for recognition, resulting in an insufficient ability to extract highly effective features. To address this limitation, in this paper, an Improved Dual-stream Shallow Convolutional Neural Network based on an Extreme Gradient Boosting Algorithm (IDSSCNN-XgBoost) is introduced for ME Recognition. The proposed method utilizes a dual-stream architecture where motion vectors (temporal features) are extracted using Optical Flow TV-L1 and amplify subtle changes (spatial features) via Eulerian Video Magnification (EVM). These features are processed by IDSSCNN, with an attention mechanism applied to refine the extracted effective features. The outputs are then fused, concatenated, and classified using the XgBoost algorithm. This comprehensive approach significantly improves recognition accuracy by leveraging the strengths of both temporal and spatial information, supported by the robust classification power of XgBoost. The proposed method is evaluated on three publicly available ME databases named Chinese Academy of Sciences Micro-expression Database (CASMEII), Spontaneous Micro-Expression Database (SMIC-HS), and Spontaneous Actions and Micro-Movements (SAMM). Experimental results indicate that the proposed model can achieve outstanding results compared to recent models. The accuracy results are 79.01%, 69.22%, and 68.99% on CASMEII, SMIC-HS, and SAMM, and the F1-score are 75.47%, 68.91%, and 63.84%, respectively. The proposed method has the advantage of operational efficiency and less computational time.

KEYWORDS

ME recognition; dual stream shallow convolutional neural network; euler video magnification; TV-L1; XgBoost

Abbreviation

ME	Micro Expressions
EVM	Euler Video Magnification
DNNs	Deep Neural Networks
IDSSCNN	Improved Dual-Stream Shallow Convolutional Neural Network

XgBoost	Extreme Gradient boosting algorithm
LBP	Local Binary Patter
DL	Deep Learning
FE	Facial Expression
TV-L1	Total Variation regularized L1-norm minimization
CBAM	Convolutional Block Attention Module
CASMEII	Chinese Academy of Sciences Micro-Expression Database
SMIC-HS	Spontaneous Micro-Expression Database
SAMM	Spontaneous Actions, and Micro-Movements
ReLU	Rectified Liner Unit
TP	True Positive
TN	True Negative
FP	False Positive
FN	False Negative
RAM	Random Access Memory
P	Precision
R	Recall

1 Introduction

1.1 Motivation

Micro-expressions (ME), which last for a second, are spontaneous expressions that reveal people's unconscious behaviour and provide insight into their emotional state. These expressions can not be faked, making them accurate indicators of an individual's genuine emotions. ME recognition has many potential applications in various fields, including clinical diagnosis, polygraph detection, education, and business. These expressions are often challenging to detect due to their short duration (0.05 to 0.2 s) and subtlety. This motivates the need for more advanced and accurate recognition systems, mainly using deep learning techniques.

1.2 Literature Review

In recent years, automatic ME recognition has attracted more and more attention. In ME broadly, two main approaches are used to extract features, namely, hand-crafted and deep learning-based methods. Hand-crafted methods capture appearance-based or geometric-based ME features by manually designed visual descriptors and then feed these features into a classifier for emotion recognition. Typical methods based on hand-craft features are LBP-TOP algorithm [1]. The first author who presented the variations of the LBP-TOP algorithm was Guo et al. [2]. He proposed a novel feature extraction approach, centralized binary patterns on three orthogonal planes (CBP-TOP), which leverages information from three orthogonal planes. Their methodology prioritizes pixels only with the maximum weightage relative to their peers. Similarly, in [3], the author introduced spatio-temporal local binary patterns with integral projection. They employed integral projection in their proposed approach to retain facial image form features and enhance ME discrimination. Moreover, Huang et al. [4] presented Spatio-Temporal Completed Local Quantized Patterns (STCLQP). Instead of homogeneous patterns, they constructed suitable pattern codes using orientation, sign, and magnitude components. These pattern-based codes can improve performance, but this depends on the training dataset used to build the classifiers.

Additionally, several researchers have investigated using the optical flow method for feature extraction. Optical flow is calculated using the brightness differences between consecutive frames, allowing for assessing small facial motions. Several optical flow-based techniques have been explored. For instance, Liu et al. [5] proposed the primary directional mean optical flow feature (MD-MO) as a key attribute in their method. In contrast, Liong et al. [6] proposed the bi-weighted oriented optical flow (Bi-WOOF) method to extract the meaningful feature for ME. Previously, the discussed method performed well in extracting meaningful features. Still, ME recognition is a subtle change in human emotions, and it's hard for the traditional method to recognize these subtle moments. To address this issue, the authors in [7] used the EVM algorithm to amplify the ME movement before identifying it using traditional recognition methods. However, handcrafted methods usually require researchers with expertise to select representative descriptors, and the parameter-tuning process is complicated. In addition, these methods have low robustness and resist adapting to the impacts of incorporating novel data.

In contrast, DL methods utilize neural networks to extract features from data automatically. DL techniques are evaluated using a variety of architectures, such as convolutional neural networks (CNN) and recurrent neural networks (RNNs). Goh et al. [8] were among the pioneers in implementing DL-based algorithms for ME recognition. In their work, CNN was employed to identify spatial features within select video frames at different levels of expression (onset, apex, and offset). Subsequently, to recognize the time-dependent features in video sequences, authors utilized the RNN-based Long Short-Term Memory (LSTM) method. Similarly, Peng et al. [9] introduced the Dual Temporal Scale convolutional Neural Network (DTSCNN). Their work incorporates optical-flow sequences across various temporal scales to measure the high-dimensional spatio-temporal feature space.

Following the work of Wang et al. [10] and Reddy et al. [11] proposed a 3D-CNN architecture, They used a 3D convolutional kernel on a video sequence for feature identification and processing. Moreover, Takalkar et al. [12] proposed a hybrid method that combined DL and handcrafted features. The LBP-TOP algorithm was used for handcrafted features, which capture spatio-temporal facial motions. In contrast, deep features are derived using CNN. The recognition process employs DL algorithms in a black box. Despite advancements in DL-based algorithms for ME recognition, their reliability remains challenging due to the limited number of ME datasets. Zhi et al. [13] and Wang et al. [14] addressed this issue by using transfer learning and fine-tuning their models on ME datasets. However, transfer learning-based methods can be affected by noise, such as brightness, misaligned faces, and the short duration, subtle movement of ME.

1.3 Contribution and Paper Organization

ME, subtle and fleeting facial expressions, hold significant potential for applications in various fields, including human-computer interaction, security, and healthcare. However, their subtle nature poses a considerable challenge for accurate and robust recognition. Existing methods often rely on deep CNN, which can be computationally expensive and prone to overfitting due to limited training data. Additionally, these methods usually overlook the importance of spatial and temporal information in ME analysis.

To address these limitations, we propose a novel approach that preprocesses the initial and keyframes of ME videos using the EVM algorithm and the TV-L1 optical flow algorithm. These algorithms enhance subtle motion signals and facilitate more accurate detection of motion flow. Subsequently, we extract spatial and temporal features from the enhanced frames, which serve as inputs to the IDSSCNN. This network is further improved by integrating an attention mechanism CBAM that selectively focuses on the most informative features, improving the model's efficiency in handling ME. Finally, we employ the XgBoost algorithm in the classification layer to boost the precision and accuracy of the ME recognition process. The key contributions of this method are highlighted below:

A. The dual-stream architecture effectively captures spatial and temporal information from ME videos, leading to more discriminative feature representations.
B. The integration of XgBoost enhances the model's ability to capture complex patterns and improve classification accuracy.
C. IDSSCNN-XgBoost uses auto-feature learning effectively and recognizes ME with higher accuracy. Experiments on various ME databases, such as CASMEII, SMIC-HS, and SAMM demonstrate the effectiveness of our proposed approach compared to existing methods [15–17]

This paper is structured as follows: Section 2 related work is discussed briefly. Section 3 provides an overview of relevant methods, including the EVM algorithm, TV-L1 optical flow algorithm, dual stream convolutional neural network, and convolutional block attention module. Section 4 introduces the proposed method. Section 5 outlines the experimental conditions and compares the results. Finally, in Section 6, we summarize the paper's findings and provide an outlook for future research.

2 Related Work

Many researchers have devoted a considerable amount of effort over a period to improve computers' ability to understand human FE and emotional actions. According to Takalkar et al. [18], numerous techniques have been developed in ME detection and recognition, including filtering and classification. In this section, we compare the research status of ME identification based on handcraft features and DL, obtain the limitations of existing research, and set the framework for the research approach.

2.1 Handcrafted Features

In ME recognition, extracting valuable features is crucial for enhancing recognition accuracy. Huang et al. [4] introduced the spatio-temporal local quantization pattern (STCLQP), which expands on LBP-TOP by improving the acquisition of input data through the integration of pixel variation in sign, magnitude, and orientation. This approach also establishes a concise spatio-temporal domain codebook to optimize recognition outcomes. Additionally, Huang et al. [3] further included facial shape characteristics in spatiotemporal texture features and proposed a spatiotemporal local binary pattern (STLBP) for ME recognition.

Researchers are employing optical flow-based approaches to extract motion-related information from ME videos or sequences for recognition. Liu et al. [19] developed the main directional mean optical flow (MDMO) approach, which uses optical flow to create a feature vector that describes the local motion of an ME on a face. This feature vector is fed into a support vector machine (SVM) for ME recognition. Based on this study, Liong et al. [20] introduced the bi-weighted oriented optical flow (Bi-WOOF) feature descriptor, which uses only the start and peak frames to characterize a delicate ME sequence. The Bi-WOOF approach uses the optical flow and strain magnitude as weights to create directional histograms of face area blocks. This highlights the significance of each optical flow for ME recognition. In addition, Ni et al. [21] developed the LGSNet, a dual-stream network that integrates optical streams and segment-level video features. The merging of information improves the identification skills of the LGSNet for ME analysis. To divide the facial region into regions of interest (ROIs) and localize facial landmarks, Li et al. [22] developed a deep multitask learning-based approach for ME recognition. Facial muscle activity produces ME, so to determine the direction of facial muscle movement, a robust optical flow technique is paired with histograms of oriented optical flow (HOOF) characteristics [23]. Afterwards, SVM is used as a classifier to identify ME. Motion recordings at several face areas, like the corners of the mouth and eyebrows, are necessary for the facial action coding system (FACS), a crucial tool for identifying ME. Using both ROIs and HOOF characteristics, this specific technique produces precise ME identification that maps to action units (AUs).

2.2 Deep Learning Methods (DL)

With the development of DL technology, a series of deep neural network structures have been applied to ME recognition. Zhou et al. [24] employed a dual-stream inception network to combine the properties of a particular ME to recognize ME better and obtain prominent and discriminative features of a given expression. To embed spatial and temporal information into ME video clips, Liong et al. [25] proposed a shallow three-stream 3D CNN method. In their work CNN learns from three optical flow features, namely, optical strain, horizontal and vertical optical flow fields, which are computed based on the start and apex frames of each video and extract discriminative high-level features and ME details through lightweight computation. Li et al. [26] proposed a deep local holistic network for ME recognition, leveraging subnetwork fusion and attention mechanisms to highlight key features and reduce background noise. Van Quang et al. [27] proposed a CapsuleNet-based ME recognition method for apex frame ME sequences, overcoming information loss caused by maxpooling in traditional image processing and enhancing recognition robustness and accuracy by preserving spatial link between features. The concept of spatio-temporal transformations was initially presented by Xia et al. [28]. They also built the spatiotemporal recurrent convolutional networks (STRCN) model, effectively combining spatial features and dynamic changes through dual-time dimension processing. It significantly improves the ability to interpret fast and subtle facial muscle movements. Su et al. [29] utilized a component-aware attention module to capture motivational information and non-rigid deformations by highlighting key ME. Gajjala et al. [30] developed a 3D residual attention network (MERANet) to recognize facial expressions and classify emotions. The model leverages the strengths of spatial-temporal and channel attention. MERANet performs well on limited datasets. However, the low sample size in ME datasets limits the model's generalizability. Zhou et al. [31] proposed the dual branch attention network (Dual-ATME) to address the issue of ineffective single-scale features in ME recognition. However, the ability to adapt and learn in complex and rapidly changing scenarios remains challenging due to the reliance on prior knowledge.

Current ME recognition algorithms have made significant progress in detecting ME and micro-features using various optimization techniques. However, the accurate measure of an ME recognition algorithm's effectiveness lies in its ability to extract robust and generalizable features that perform well across different scenarios. This paper proposes a method that enhances the network's ability to effectively capture temporal and spatial features, outperforming other neural networks. Additionally, the shallow architecture of our network allows it to be trained efficiently on small datasets, enabling end-to-end learning without the need for extensive data resources.

3 Materials and Methods

3.1 Euler Video Magnification Algorithm (EVM)

In 2012, Wu et al. [32] and Liu et al. [33] developed the EVM algorithm based on fluid Euler equations. This algorithm analyzes the relationship between pixels and time by considering the pixels in a video as a function of space and time. The EVM algorithm consists of three main steps: spatial filtering, temporal filtering, and image amplification using composite images. During the spatial filtering, noise and unwanted details are removed from the video frames. The temporal filtering step analyzes the changes in pixel values over time to identify subtle movements that are not visible to the naked eye. Finally, image amplification is achieved using composite images highlighting the identified movements. The EVM algorithm uses a multi-resolution image pyramid to break the video into different spatial frequencies. By applying temporal filtering to each scaled image, it is possible to isolate the frequency bands of interest. Generally, facial expressions are characterized by a concentration of movements in the low-frequency range, with only a few features, such as the eyebrow, nose, and lip motions, remaining unchanged. Once the Taylor series approximates the filtered frequency band, it is amplified by an "e" factor to enhance the filtering result. Finally, the original and amplified signals are combined to create a composite image, as shown in Fig. 1.

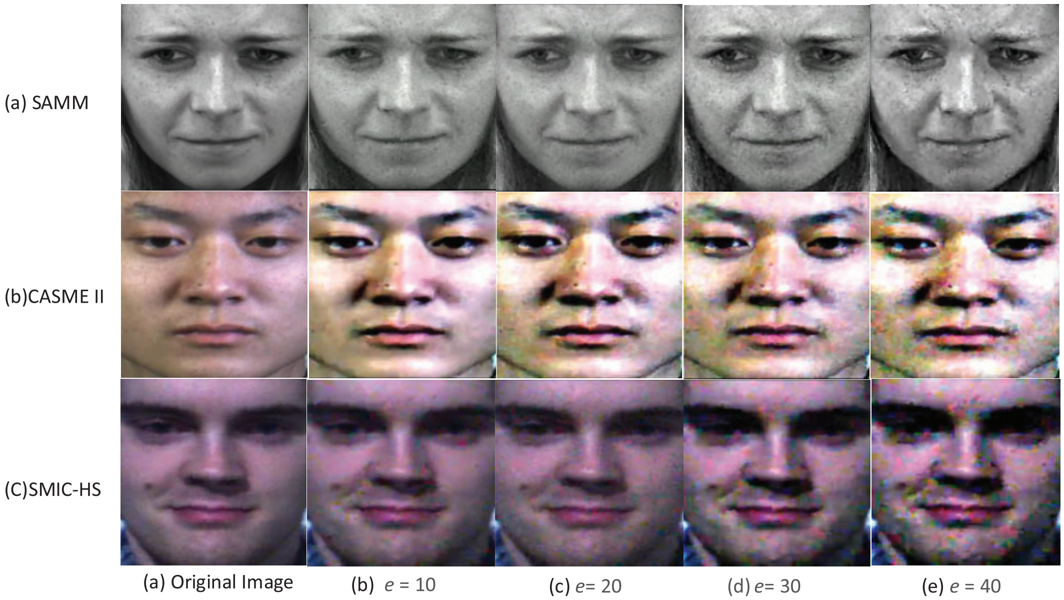

Figure 1: Comparison of images at different magnification factors. (a) image was taken from the SAMM dataset [34] and reproduced here following the SAMM license agreement. Copyright ©Yap: fuxl@psych.ac.cn. Similarly, the (b) images are from subject 1 of the CASME II dataset [15] and is reproduced here following the CASME2 license agreement. Copyright ©Xiaolan Fu, (c) The SMIC-HS dataset was provided with permission by license agreement. Copyright ©Li Xiaobai research group at the University of Oulu Finland [35]

3.2 Total Variation Regularized L1-Norm Minimization (TV-L1)

The optical flow method is a widely used technique for analyzing the movement of objects, including the subtle movements of facial muscles. By generating an optical flow image, researchers can gain insights into the dynamics of a given scene [19,36,37]. Additionally, motion analysis can help preserve important edge feature information in the image, allowing for more accurate analysis of the movement patterns [38].

Suppose there are two frames, I_0 and I_1, and the pixel value, $X = (x, y)$, is located on frame I_0, as shown in Eq. (1), where μ is the scaling factor that adjusts the influence of the gradient term ∇I_1.

$$I'_1 = \mu \nabla I_1 + I_1(X + U_0) - U_0 \nabla I_1 - I_0 \tag{1}$$

Eq. (2) depicted the energy function of the TV-L1 method. $U = (u, v)$ represents the optical flow field in u and v directions, and Ω is the spatial domain over which the optical flow is computed. Similarly, λ is the regularization parameter that controls the balance between the data fidelity $\lambda\,|I'_1|$ and total variation regularization term $\nabla\,|U|$. On the other hand, ∇u and ∇v are the second derivatives of u and v. The first item in Eq. (1) represents the gray value difference of the corresponding pixels in the adjacent frames, and the second term represents the regularization constraint for the motion.

$$E = \int_{\Omega} \{\lambda\,|I'_1| + |\nabla U|\}\,dx \tag{2}$$

The TV-L1 algorithm employs a dual approach to reduce the optical flow energy function. Firstly, it used a Taylor expansion technique to approximate the grayscale difference of adjacent frames. Secondly, it used a bidirectional method to update the variable alternatively to reduce the error as shown in Eq. (3):

$$I''_1 = U \nabla I_1 + I_1(X + U_0) - U_0 - I_0$$

$$E = \int_{\Omega} \{\lambda\,|I''_1| + |\nabla U|\}\,dx \tag{3}$$

To measure the discrepancy between the predicted optical flow and the observed image gradients, the fidelity term is defined as $\rho U = I_1(X + U_0) + (U - U_0)\nabla I_1 - I_0$ while subtracting the U' from the Eq. (3) yields Eq. (4):

$$E = \int_{\Omega} \left\{\lambda\,|\nabla U| + \frac{1}{2\theta}(U - U')^2 + \rho U\right\} dx \tag{4}$$

In Eq. (4), θ is a constant, and U' approximates U as the number of iterations increases. The formula optimizes the below function by iteratively updating U' and U. Their threshold functions are shown in Eq. (5):

$$U' = U + \begin{cases} \lambda\theta I_1^X(U), \rho U < -\lambda\theta (I_1^X)^2 \\ -\lambda\theta I_1^X(U), \rho U > -\lambda\theta (I_1^X)^2 \\ {}^{-\rho U}/_{I_1^X}, |\rho U| \le \lambda\theta (I_1^X)^2 \end{cases} \tag{5}$$

The TV-L1 provides significant advantages in computing the ME features. During the TV-L1 method to protect the edges from blurring during diffusion, the TV-L1 method maintains regularization term displacement field discontinuities. In this way, we can preserve the edge information for ME during diffusion. Furthermore, the data term employs a strong L1 norm to tolerate brightness changes. We created optical flow feature maps for various basic ME images, and the results accurately depicted ME movements.

3.3 Dual Stream Convolutional Neural Network

In the field of video recognition, it is necessary to capture information that changes over time between images. Simonyan et al. [39] presented a dual-stream convolutional network for video action recognition to exploit this data efficiently. This network employs two independent convolutional networks to extract spatial and temporal features from images before combining the classification results of both networks to produce the final recognition result. Temporal convolutional networks were trained using optical flow maps obtained between frames, and spatial convolutional networks were learned using video frames.

3.4 Convolutional Block Attention Module (CBAM)

The Convolutional Block Attention Module (CBAM) is a novel attention mechanism module that has been proposed by Woo et al. [40] to enhance the performance of convolutional neural networks (CNNs). It creates attention maps in channel and spatial dimensions to capture essential features effectively. The attention maps generated by CBAM are used to weight the feature maps, thereby increasing the network's ability to focus on crucial input regions. The entire structure of the attention model is depicted in Fig. 2, which shows the flow of information through the network. The attention operation process can be expressed using the mathematical formula below that captures the essence of the CBAM module.

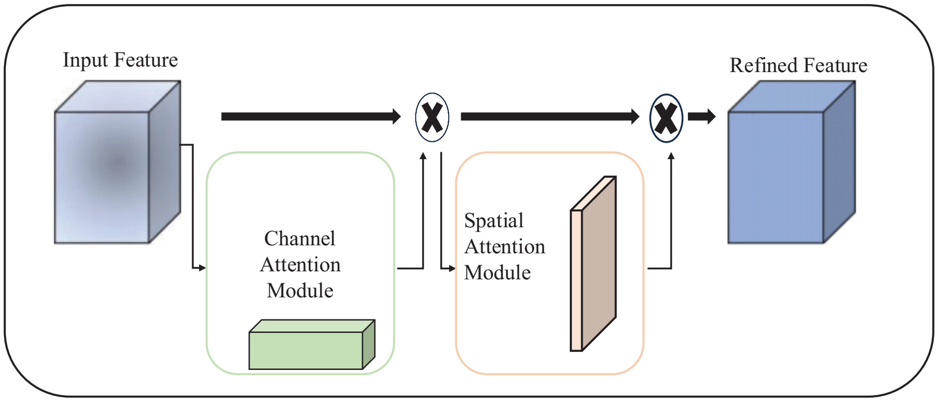

Figure 2: The overview of CBAM

In Eq. (6), $F \in R^{C\times H\times W}$ represents the input feature (i.e., C represents the number of channels, H represents the height, and W the width) map before entering the CBAM. Similarly, $M_c \in R^{c\times 1\times 1}$ represents the channel attention map $M_s \in R^{1\times H\times W}$ is the spatial attention map, and $\otimes$ represents the multiplication of corresponding elements. F is then processed by the channel attention module and spatial attention module as follows:

$$F' = M_c(F) \otimes F$$

$$F'' = M_s(F') \otimes F' \tag{6}$$

where F' represents the result of the feature map multiplying the channel map, and F'' represents the result of the spatial attention map multiplying F' or the final output.

The CBAM module comprises two sub-modules, namely channel and spatial. Its purpose is to adaptively refine the intermediate feature map at every convolutional block of deep networks. The channel attention map capitalizes on the inter-channel relationship of features, considering each channel of a feature map as a feature detector. This enables the attention mechanism to concentrate on what is pertinent in an input image.

The shared layer in Fig. 3 comprises a multi-layer perceptron (MLP) with one hidden layer. To simplify the calculation process, the activation size of the hidden layer is set to $R^{c/a \times 1 \times 1}$, where a is the reduction rate. The specific calculation for channel attention is given in Eq. (7):

$$F_1 = MLP(\mathrm{F}_{avg}^{c}) = W_1(W_0(\mathrm{F}_{avg}^{c}))$$
$$F_2 = MLP(\mathrm{F}_{max}^{c}) = W_1(W_0(\mathrm{F}_{max}^{c}))$$
$$M_c(F) = \sigma(F1 + F2) \tag{7}$$

In Eq. (7), σ is the sigmoid function, and $W_0 \in R^{c/a \times c}$, $W_1 \in R^{c \times c/a}$, F_{avg}^{c}, and $F_{\max}^{c}$ represent the average pooling feature and maximum pooling feature. The ReLU activation function is used after the MLP weight W_0.

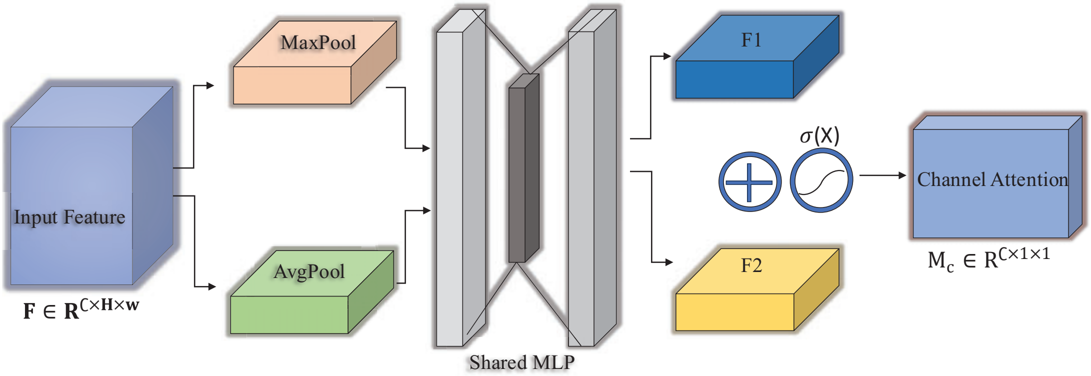
MaxPool
AvgPool
Input Feature
$\mathbf{F} \in \mathbf{R}^{C\times H\times w}$
Shared MLP
F1
F2
$\sigma(X)$
Channel Attention
$M_c \in R^{C\times 1\times 1}$

Figure 3: Channel attention module

The spatial attention model is calculated as Eq. (8):

$$M_s(F) = \sigma\left(f^{7\times 7}\left(\left[F_{avg}^{s}; F_{\max}^{s}\right]\right)\right) \tag{8}$$

In Eq. (8), σ represents the sigmoid function, and $F_{avg}^{s} \varepsilon R^{1\times H\times W}$ and $F_{max}^{s} \varepsilon R^{1\times H\times W}$ represent the average feature fusion and the maximum fusion feature on the channel, which produce feature map with dimensions $(1 \times H \times W)$, are merged into a single feature map with dimensions $(2 \times H \times W)$. This combined feature map is then processed through a $f^{7\times 7}$ convolution operation, reducing the number of channels back to 1 $(1 \times H \times W)$ while capturing spatial relationships within the image. $f^{7\times 7}$ also provides a balance between capturing sufficient contextual information and keeping the computational cost reasonable. The specific operation of the spatial attention model is shown in Fig. 4.

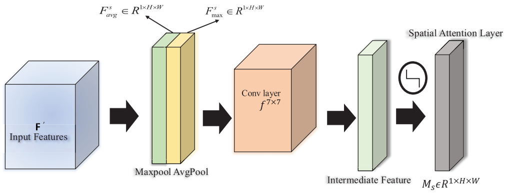
$F_{avg}^{s} \in R^{1\times H\times W}$
$F_{\max}^{s} \in R^{1\times H\times W}$
Spatial Attention Layer
F′
Input Features
Maxpool AvgPool
Conv layer $f^{7\times 7}$
Intermediate Feature
$M_S \epsilon R^{1\times H\times W}$

Figure 4: Spatial attention module

3.5 Extreme Gradient Boosting (XgBoost)

Extreme Gradient Boosting (XgBoost) [41] is a widely utilized tree-based machine learning technique for data classification, owing to its efficiency. Unlike Gradient Boosting Machines (GBC) [42], which construct each ensemble tree sequentially, XgBoost parallelizes the job, resulting in considerable speed gains. Fig. 5 depicts XgBoost as an additive model that linearly combines multiple decision trees. In XgBoost, each subsequent tree is constructed sequentially based on the results of the preceding tree until the final tree is formed.

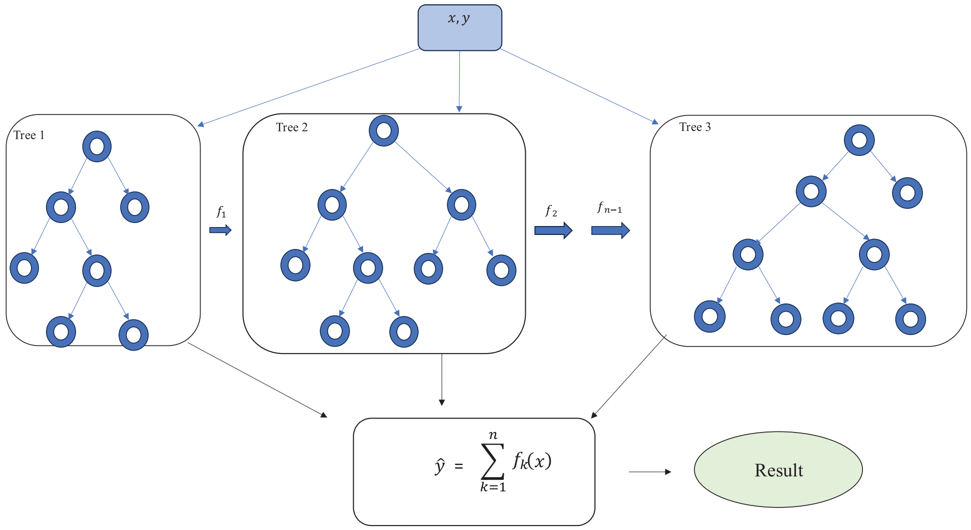

Figure 5: General architecture of the XgBoost algorithm

Furthermore, the weight of each tree changes after each stage of the iterative boosting technique, which focuses on fitting residuals. In this context, residuals are the discrepancy between the previous tree's actual and expected values. This process is repeated until a residual value of zero is obtained. Ultimately, the predictions from all the trees are combined to formulate the final prediction of the XgBoost model. The objective function of the XgBoost model comprises two terms. The loss function is the first term, while the regularization term is the second. The loss function can be used to calculate the difference between the predicted and actual value, and the regularization terms can be utilized to avoid overfitting.

4 Proposed Method

In video recognition, capturing information that changes over time between different frames is necessary. The authors of Xie et al. [43] and Tang et al. [44] proposed a dual-stream convolutional network for motion recognition. Inspired by their work in this paper, we proposed the IDSSCNN-XgBoost algorithm. Fig. 6 depicts the IDSSCNN-XgBoost structure. The Proposed method performed the following steps: preprocessing, feature extraction, and ME recognition and classification.

In the preprocessing step, face alignment and face cropping of all image frames between the ME Onset and Apexset frames are performed. Then, optical flow TV-L1 and EVM methods are utilized in parallel to extract temporal and spatial features. These features are then processed by IDSSCNN, with the help of the CBAM attention module, to refine the extracted effective features. The outputs are then fused, concatenated, and classified using the XgBoost algorithm.

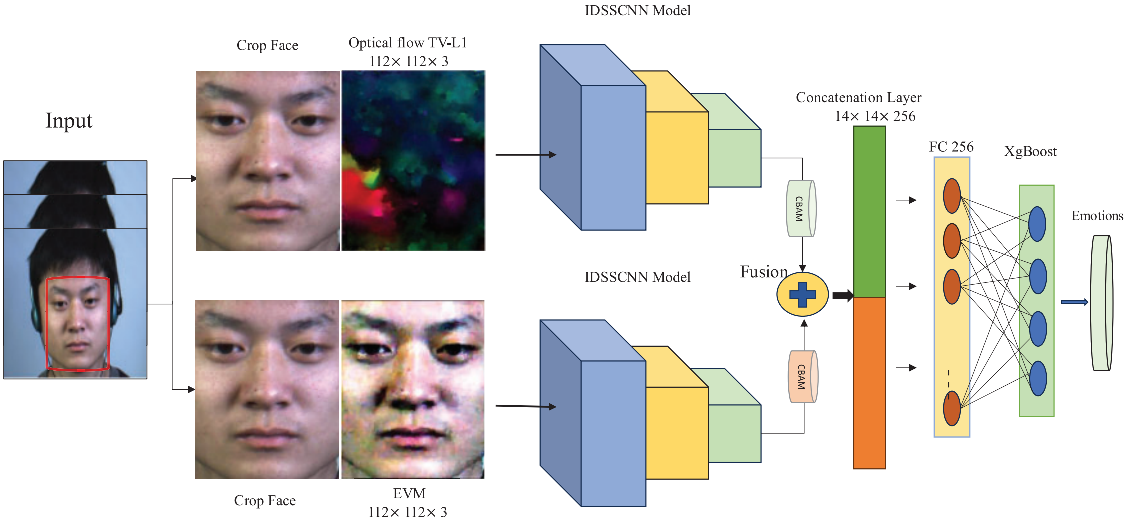

Figure 6: The proposed IDSSCNN-XgBoost algorithm

4.1 *Preprocessing*

A. Face detection and cropping
We perform face location detection and cropping to reduce the influence of irrelevant backgrounds on face images. Firstly, we used the dlib library [45] on each frame to locate feature points on the face and then used these key points to split the rectangle recognition frame of the face area. Secondly, because the face image has multiple postures and angles, using the dlib library, face images are normalized. Finally, we got the facial image with the size 112 × 112.

B. EVM
EVM amplification is employed to extract and amplify the spatial feature, as discussed in Section 3.1. Time domain filtering is performed using the Butterworth bandpass filter, with an amplification frequency range of 1–5 Hz. During the signal amplification step, we compare the effect of different amplification factors on the image. Our findings show that a magnification factor e of 10 does not produce noticeable image magnification, and facial movements are not visible. On the other hand, a magnification factor greater than 20 increases the influence of noise in the image, which may cause facial expression distortion. These results suggest an optimal magnification factor should be selected to ensure accurate facial expression recognition.

C. Optical flow TV-L1 method
The TV-L1 optical method is reported to be one of the most efficient approaches for ME recognition, which is outlined in Section 3.2. This paper used the TV-L1 method to extract optical flow temporal features from ME video sequences to generate an optical flow dataset.

4.2 *IDSSCNN Network Design*

Table 1 presents the architecture of IDSSCNN-XgBoost. In this setup, a magnified image from EVM is fed into one block, while optical flow TV-L1 features are fed into a second block. Each independent IDSSCNN is responsible for extracting temporal and spatial features. The proposed method utilizes dilated convolutional layers and attention mechanisms to enhance the model's capability to capture high-dimensional features. Dilated convolutional is a convolutional operation used in CNN to enhance the receptive area of the network without increasing the number of parameters. The average fusion layer F^{s}_{avg} and maximum F^{s}_{max} fusion layer then fuse the two layers convolutional responses to reduce the loss of delicate image features in the convolutional layer. Unlike traditional neural networks, in the proposed method, a pooling layer is not added because it prevents the loss of minor movements of ME. Instead of a pooling layer, a convolution with a stride size of 2 is used.

Table 1: IDSSCNN-XgBoost network architecture

Layer name	Input size	Kernel size	Stride	Output size
Original image	112 × 112 × 3	–	–	112 × 112 × 3
Convolutional layer (branch 1)	112 × 112 × 3	1 × 1	2	56 × 56 × 16
Convolutional layer (branch 2)	112 × 112 × 3	1 × 1	2	56 × 56 × 16
Fusion layer	56 × 56 × 16	–	–	56 × 56 × 16
Attention module	56 × 56 × 16	–	–	56 × 56 × 16
Dilated convolutional layer	56 × 56 × 16	3 × 3	2	28 × 28 × 32
Dilated convolutional layer (branch 1)	28 × 28 × 32	3 × 3	2	28 × 28 × 32
Dilated convolutional layer (branch 2)	28 × 28 × 32	3 × 3	2	28 × 28 × 32
Attention module	28 × 28 × 32	–	–	28 × 28 × 32
Fusion layer	28 × 28 × 32	–	–	28 × 28 × 32
Dilated convolutional layer	28 × 28 × 32	5 × 5	2	14 × 14 × 64
Dilated convolutional layer	28 × 28 × 32	5 × 5	2	14 × 14 × 64
Concatenation layer	14 × 14 × 256	–	–	14 × 14 × 256
Convolutional layer	14 × 14 × 256	7 × 7	–	7 × 7 × 256
Fully connected layer	7 × 7 × 256	–	–	256
XgBoost	256	–	–	–

The parallel processing of spatial and temporal aspects helps enhance the network's ability to recognize and interpret MEs accurately, improving overall performance in this context. While the network effectively extracts essential features, its recognition accuracy is limited due to softmax, resulting in a polarized output that gives one neuron a high score while giving the rest low scores. Algorithm learning is a superior approach; the recognition accuracy is high, overfitting is complex, and it has good generalization ability. However, it is difficult to achieve sound learning effects when the image features are complex and distorted.

To overcome this issue, instead of using a softmax layer for classification, the concatenated features are fed into an eXtreme Gradient Boosting (XgBoost) classifier, The XgBoost algorithm is chosen for its high operational efficiency and precision in ME recognition. This approach allows mutual learning between the two methods, improving recognition accuracy. Because of its high operational efficiency and precision, the XgBoost algorithm is a valuable tool in ME recognition. Moreover, for high-accuracy applications, where accuracy is paramount, TV-L1 and EVM, in parallel, can provide a highly cost-effective solution by enhancing both visibility and accuracy.

5 Experiment and Result Analysis

5.1 Dataset Introduction

To verify the effectiveness of the IDSSCNN-XgBoost algorithm, three public datasets, CASMEII, SMIC-HS, and SAMM, were selected. The CASMEII dataset includes 247 ME video clips from 26 different people captured under constant and high-intensity illumination conditions. The samples are divided into five categories: happiness, surprise, disgust, suppression, and others. Similarly, the SMIC-HS dataset, comprised of 164 spontaneous MEs from 16 individuals, was captured using a variety of cameras. Some samples were taken using a 100-fps high-speed camera, while others were captured using a 25-fps visible-light camera and a 25-fps near-infrared camera. The ME is categorized as positive, negative, or surprise, providing researchers with various emotional expressions to analyze. Table 2 presents the dataset's basic information. Moreover, the SAMM dataset has 159 video samples from 29 different people, with the same number of samples for three and five classes.

Table 2: Basic information of datasets used in the experiment

	CASMEII	SMIC-HS	SAMM
Frame rate (fps)	200	100	200
Expression category	7	3	8
Total number of videos	255	164	159
Happy	32	–	26
Surprise	28	43	15
Anger	–	–	57
Disgust	63	–	9
Sad	4	–	6
Fear	2	–	6
Contempt	–	–	8
Repression	27	–	12
Others	99	–	–
Positive	–	51	26
Negative	–	70	–
Start frame	Marked	Marked	Marked
Key frame	Marked	Unmarked	Marked
End frame	Marked	Marked	Marked

5.1.1 Performance Metrics

To evaluate the IDSSCNN-XgBoost model's performance, accuracy and F1-score were used as ME recognition performance metrics. By designating a particular subset of participants as the test group and utilizing the remaining participants as the training group, we apportioned the ME recognition accuracy of all participants equally.

Accuracy:

Accuracy measures the proportion of correctly classified samples in a given test dataset to the total number of samples. It is expressed through Eq. (9):

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{9}$$

TP is the abbreviation for true positive, referring to the accurate prediction of genuine emotions as positive. TN, on the other hand, denotes true negative, which signifies the correct classification of irrelevant emotions as negative emotions. FP, or false positive, refers to incorrectly predicting irrelevant emotions as positive emotions. Lastly, FN stands for false negatives and refers to the inaccurate prediction of real emotions as negative emotions.

F1-score:

F1-score is a metric that joins accuracy and recall in a single value, which is calculated as a weighted average. The formula for the F1-score is given as (10), where P stands for precision, and R signifies recall, with both carrying equal weight.

$$\text{F1-score} = 2 \times \frac{\text{P} \times \text{R}}{\text{P} + \text{R}} \tag{10}$$

5.1.2 Experimental and Parameters Setting

The proposed technique was implemented using Python 3.7 and Keras version 2.11.0 RC3 [46], a high-level open-source neural network toolkit that operates on top of TensorFlow [47]. We chose Keras for its flexibility, speed, and ease of use in experimentation. In addition, we utilized Imblearn [48] and scikit-learn [49], two Python libraries, to aid in the implementation of the technique. The simulation was carried out on a high-performance system with an Intel(R) Core(TM) i7-8750H CPU and 32 GB of RAM, running a 64-bit operating system. We used an NVIDIA GeForce GTX 1070 to train the model with a MAX-Q design GPU. To tune the proposed method, input images are reduced to 112 × 112, and an image with a magnification of 20 is selected. The proposed network employs stochastic gradient descent (SGD) for learning the network parameters [50]. The momentum, weight decay, and stopping criterion are set to 0.9, 0.0005 and 10^{-4}. Initially, the learning rate is 10^{-4}, and it will be modified in the subsequent iterations using the damping factor of 0.8. The existing ME datasets contain imbalanced classes, and even classes have a minimal number of samples. It can lead to poor performance due to overfitting. To mitigate this issue, L2 regularization is used. Furthermore, for Xgboost max_depth size was selected 3, n_estimators 500, and learnig_rate was set to 0.001, which boosts the XgBoost algorithm as shown in Table 3.

Table 3: Parameter setting of IDSSCNN-XgBoost

Parameter	Range
Image input	112 × 112
Momentum	0.9
Weight decay	0.0005
Stopping criterion	10^{-4}
Number of epochs	200
Drop out layer	0.5
Batch size	64

(Continued)

Table 3 (continued)

Parameter	Range
Max_depth size	3
n_estimators	500
Learning_rate	0.001

5.2 Result Analysis

5.2.1 Comparison with Previous Methods

Table 4 shows that the accuracy of the LBP-TOP algorithm is 39.68% on CASMEII, 43.73% on SMIC-HS, and 35.56% on SAMM, which is 18% lower than the Bi-WOOF on CASME11, 18% lower on SMIC-HS and 16% lower on SAMM. Compared to the proposed method, the IDSSCNN-XgBoost algorithm outperformed the LBP-TOP and Bi-WOOF algorithms on these publicly available datasets. The IDSSCNN-XgBoost accuracy at CASMEII was calculated to be 79.01%, which is 39.33% higher than the LBP-TOP method and 21.12% higher than the Bi-WOOF algorithm. Similarly, the proposed algorithm at SMIC-HS has an accuracy of 69.22%, which is 25.49% and 7.63% higher than the LBP-TOP and Bi-WOOF algorithms, respectively. However, for SAMM data sets, the proposed method shows an accuracy of 68.99%, which is 17.6% higher than Bi-WOOF and 33.43% higher than the LBP-TOP algorithm. VGG16 is a classic DL algorithm that is often used as a base algorithm for DL comparisons. Compared to the proposed IDSSCNN-XgBoost method, the detection accuracy improves by 8.01%, 9.58% and 21.06% for CASMEII, SMIC-HS, and SAMM, respectively. We also compare the proposed method to the DSSN algorithm, which uses a multichannel convolutional neural network structure similar to the dual-stream structure proposed in this study. Compared to the DSSN algorithm, our proposed method improved detection accuracy by 8.23% on CASMEII, 6.81% on SMIC-HC, and 11.64% on SAMM datasets.

Table 4: Comparison of the IDSSCNN-XgBoost algorithm with previous methods based on accuracy

Models	CASMEII	SMIC-HS	SAMM
LBP-TOP	0.3968	0.4373	0.3556
Bi-WOOF	0.5789	0.6159	0.5139
VGG16	0.7100	0.5964	0.4793
GoogleNet	0.6414	0.5511	0.5992
AlexNet	0.7415	0.6373	0.6643
SSSN	0.7119	0.6329	0.5662
DSSN	0.7078	0.6241	0.5735
CapsuleNet	0.7018	0.5877	0.5989
GACN	0.7120	0.6120	0.5231
KFC	0.7276	0.6585	0.6324
LGCon	0.6214	0.6341	0.3529
CDSCCN	0.7434	0.6712	0.6515
IDSSCNN-XgBoost	**0.7901**	**0.6922**	**0.6899**

According to Table 5, AexNet has the highest recognition accuracy compared to other algorithms. However, our proposed method has outperformed AlexNet regarding recognition accuracy and F1-score. Similarly, when the IDSSCNN-XgBoost algorithm was compared to the key facial components (KFC) [29], algorithm in terms of accuracy and F1-score, we can observe that while the KFC methods perform well on these three datasets, our method outperforms them significantly. IDSSCNN-XgBoost obtained 6.25% higher accuracy on the CASMEII dataset, 3.37% on the SMIC-HS dataset, and 5.75% on the SAMM dataset. Similarly, when we compare the accuracy of the proposed technique to local and global information learning (LGCon) [51] and Combining Dual-Stream Convolution and Capsule Network (CDSCCN) [52], we notice an improvement in both metrics. The results reveal that IDSSCNN-XgBoost achieved higher recognition accuracy of 79.01% on CASMEII, 69.22% on SMIC-HC, and 68.99% on SAMM datasets, which demonstrate that the proposed method can recognize the ME more efficiently and robustly than the other methods.

Table 5: Comparison of IDSSCNN-XgBoost algorithm with previous methods based on F1-score

Models	CASMEII	SMIC-HS	SAMM
LBP-TOP	0.5100	0.6025	0.3640
Bi-WOOF	0.6125	0.6110	0.3970
VGG16	0.4862	0.5025	0.2911
GoogleNet	0.5367	0.5123	0.4921
AlexNet	0.6601	0.6013	0.4200
SSSN	0.7100	0.6329	0.4500
DSSN	0.7300	0.6461	0.4640
CapsuleNet	0.7068	0.5820	0.5909
GACN	0.7225	0.6023	0.5210
KFC	0.7375	0.6638	0.5709
LGCon	0.6000	0.6200	0.2300
CDSCCN	0.7328	0.6647	0.6322
IDSSCNN-XgBoost	**0.7547**	**0.6891**	**0.6384**

The proposed method for ME recognition has been evaluated based on the F1-score. The results in Table 5 indicate that the F1-score of the IDSSCNN-XgBoost is higher than that of the traditional algorithm Bi-WOOF on the CASMEII, SMIC-HS, and SAMM databases. Moreover, the F1-score of IDSSCNN-XgBoost on the CASMEII dataset is higher than VGG16, GoogleNet, and AlexNet. Additionally, when compared with other new methods published after 2019 for ME recognition, such as SSSN, KFC, LGCon, CDSCNN, DSSN, and CapsuleNet, the IDSSCNN-XgBoost and GACN achieved better results.

5.2.2 Ablation Study

To prove the effectiveness and superiority of the IDSSCNN-XgBoost algorithm, ablation experiments were conducted with and without XgBoost, considering both EVM and TV-L1 scenarios, respectively. As shown in the Table 6, we compare the results on the three datasets. From the Table 6, we can see with XgBoost, accuracy is improved. However, without XgBoost, hardly any improvement was observed. Hence, we can confirm that integrating the XgBoost algorithm with an IDSSCNN can get higher accuracy.

Table 6: Ablation study of the proposed method

Models	CASMEII	SMIC-HS	SAMM
Without XgBoost	0.7100	0.6325	0.5640
With XgBoost	**0.7547**	**0.6891**	**0.6384**

5.2.3 Computational Complexity

In this section, we will compare the computational complexity and computation time of the IDSSCNN-XgBoost to some state-of-the-art methods. Table 7 shows each network's total parameters. The IDSSCNN-XgBoost model comprises 1.10 million learnable parameters, which is significantly smaller than previous benchmark models such as VGG16 (138M), GoogleNet (4M), SqueezeNet (1.24M), and ResNet (11M). Furthermore, the IDSSCNN-XgBoost architecture has fewer depth channels and hidden layers than previous approaches. IDSSCNN-XgBoost has 14 layers. In comparison, VGG16, GoogleNet, SqueezeNet, and ResNet have 16, 22, 18, and 34 layers.

Table 7: Comparison of the computation time of IDSSN-XgBoot with some state-of-the-art models

Models	Layer	Parameters in (millions)
ResNet	34	11
VGG16	16	138
GoogleNet	22	4
SqueezeNet	18	1.24
DSSCNN-XgBoost	**14**	**1.10**

Similarly, Table 8 presents the comparison of IDSSCNN-XgBoost with some of the most advanced models available, including factors such as image input size and computation time, which denote the testing time for one subject, that are influenced by the network's depth and complexity. The IDSSCNN-XgBoost in this study surpasses AlexNet, GoogleNet, and VGG16 in-depth and model complexity, resulting in a longer testing time for each sample. Additionally, the network designed in this study integrates the features of ME, allowing it to outperform traditional DNNs in terms of computation time and accuracy.

Table 8: Comparison of the computation time of IDSSN-XgBoot with some state-of-the-art models

Models	Image input size	Execution time (s)
AlexNet	227 × 227 × 3	18.9007
VGG16	224 × 224 × 3	95.4436
GoogleNet	224 × 224 × 3	25.4002
IDSSCNN-XgBoost	**112 × 112 × 3**	**14.4017**

6 Conclusion

This study presents a new, improved dual-stream shallow convolutional neural network based on an extreme gradient boosting algorithm (IDSSCNN-XgBoost). The proposed method uses a stride size 2 convolutional layer instead of a pooling layer. Using a dual-stream approach, the presented method effectively integrates temporal and spatial feature extraction. Temporal features are captured via TV-L1, and spatial features are enhanced using EVM. Separate IDSSCNNs process these features. The CBAM is employed to refine these features before fusion. The fused features are concatenated and classified using XgBoost. The experiments are conducted using the publicly accessible dataset CASME II, SMIC-HS, SAMM, and the findings reveal that the modifications made in this experiment improve ME recognition accuracy. The proposed method significantly improves ME recognition accuracy by leveraging the complementary strengths of both temporal and spatial information and the robust classification power of XgBoost. The effectiveness of the proposed IDSSCNN-XgBoost in ME recognition is impacted by feature engineering, model complexity, hyperparameter tuning, data imbalance, and generalizability. It is also important to tackle class imbalance and optimize hyperparameters. Future research will concentrate on deeper architectures, developing strong feature representations, and implementing advanced optimization techniques to improve the model's performance further.

Acknowledgement: The authors are thankful to Development Program of Jiangsu Province and CTBC Bank Taiwan.

Funding Statement: This work is supported by the Key Research and Development Program of Jiangsu Province under Grant BE2022059-3, CTBC Bank through the Industry-Academia Cooperation Project, as well as by the Ministry of Science and Technology of Taiwan through Grants MOST-108-2218-E-002-055, MOST-109-2223-E-009-002-MY3, MOST-109-2218-E-009-025, and MOST431 109-2218-E-002-015.

Author Contributions: Adnan Ahmad: conceptualization, data curation, formal analysis and writing the original draft; Irfan Tariq: experimental support and discussion; Zhangran He: experimental support, formal analysis and editing; Zhao Li: funding acquisition, project administration, supervision, conceptualization, writing—review and editing. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: To validate the model, we used the following public datasets. https://helward.mmu.ac.uk/STAFF/M.Yap/dataset.php (accessed on 20 April 2024), http://casme.psych.ac.cn/casme/e1 (accessed on 17 March 2024), http://www.cse.oulu.fi/SMICDatabase (accessed on 17 March 2024).

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

References

[1] G. Zhao and M. Pietikainen, "Dynamic texture recognition using local binary patterns with an application to facial expressions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 6, pp. 915–928, 2007. doi: 10.1109/TPAMI.2007.1110.

[2] Y. Guo, C. Xue, Y. Wang, and M. Yu, "Micro-expression recognition based on CBP-TOP feature with ELM," *Optik*, vol. 126, no. 23, pp. 4446–4451, 2015. doi: 10.1016/j.ijleo.2015.08.167.

[3] X. Huang, S. -J. Wang, G. Zhao, and M. Piteikainen, "Facial micro-expression recognition using spatiotemporal local binary pattern with integral projection," in *IEEE Int. Conf. Comp. Vis.*, Santiago, Chile, Dec. 7–13, 2015, pp. 1–9.

[4] X. Huang, G. Zhao, X. Hong, W. Zheng, and M. Pietikäinen, "Spontaneous facial micro-expression analysis using spatiotemporal completed local quantized patterns," *Neurocomputing*, vol. 175, pp. 564–578, 2016. doi: 10.1016/j.neucom.2015.10.096.

[5] Y. -J. Liu, J. -K. Zhang, W. -J. Yan, S. -J. Wang, G. Zhao and X. Fu, "A main directional mean optical flow feature for spontaneous micro-expression recognition," *IEEE Trans. Affect. Comput.*, vol. 7, no. 4, pp. 299–310, 2015. doi: 10.1109/TAFFC.2015.2485205.

[6] S. -T. Liong, J. See, K. Wong, and R. C. -W. Phan, "Less is more: Micro-expression recognition from video using apex frame," *Signal Process.: Image Commun.*, vol. 62, pp. 82–92, 2018. doi: 10.1016/j.image.2017.11.006.

[7] H. I. Shahadi, J. Zaid, Al-Allaq, H. J. Albattat, and Shahadi, "Efficient denoising approach based eulerian video magnification for colour and motion variations," *Int. J. Electric. Comput. Eng.*, vol. 10, no. 5, pp. 4701–4711, 2020.

[8] K. M. Goh, C. H. Ng, L. L. Lim, and U. U. Sheikh, "Micro-expression recognition: An updated review of current trends, challenges and solutions," *Vis. Comput.*, vol. 36, no. 3, pp. 445–468, 2020. doi: 10.1007/s00371-018-1607-6.

[9] M. Peng, C. Wang, T. Chen, G. Liu, and X. Fu, "Dual temporal scale convolutional neural network for micro-expression recognition," *Front. Psychol.*, vol. 8, 2017, Art. no. 1745. doi: 10.3389/fpsyg.2017.01745.

[10] S. -J. Wang *et al.*, "Micro-expression recognition with small sample size by transferring long-term convolutional neural network," *Neurocomputing*, vol. 312, pp. 251–262, 2018. doi: 10.1016/j.neucom.2018.05.107.

[11] S. P. T. Reddy, S. T. Karri, S. R. Dubey, and S. Mukherjee, "Spontaneous facial micro-expression recognition using 3D spatiotemporal convolutional neural networks," in *IEEE. Int.Conf. Neul. Net.*, Budapest, Hungary, Jul. 14–19, 2019, pp. 1–8.

[12] M. A. Takalkar, M. Xu, and Z. Chaczko, "Manifold feature integration for micro-expression recognition," *Multimed. Syst.*, vol. 26, pp. 535–551, 2020. doi: 10.1007/s00530-020-00663-8.

[13] R. Zhi, H. Xu, M. Wan, and T. Li, "Combining 3D convolutional neural networks with transfer learning by supervised pre-training for facial micro-expression recognition," *IEICE Trans. Inf. Syst.*, vol. 102, no. 5, pp. 1054–1064, 2019. doi: 10.1587/transinf.2018EDP7153.

[14] C. Wang, M. Peng, T. Bi, and T. Chen, "Micro-attention for micro-expression recognition," *Neurocomputing*, vol. 410, pp. 354–362, 2020. doi: 10.1016/j.neucom.2020.06.005.

[15] W. J. Yan *et al.*, "CASME II: An improved spontaneous micro-expression database and the baseline evaluation," *PLoS One*, vol. 9, no. 1, 2014, Art. no. e86041. doi: 10.1371/journal.pone.0086041.

[16] S. Thuseethan, S. Rajasegarar, and J. Yearwood, "Deep3DCANN: A deep 3DCNN-ANN framework for spontaneous micro-expression recognition," *Inf. Sci.*, vol. 630, pp. 341–355, 2023. doi: 10.1016/j.ins.2022.11.113.

[17] A. K. Davison, C. Lansley, N. Costen, K. Tan, and M. H. Yap, "Samm: A spontaneous micro-facial movement dataset," *IEEE Trans. Affect. Comput.*, vol. 9, no. 99, pp. 116–129, 2018. doi: 10.1109/TAFFC.2016.2573832.

[18] M. Takalkar, M. Xu, Q. Wu, and Z. Chaczko, "A survey: Facial micro-expression recognition," *Multimed. Tools Appl.*, vol. 77, no. 15, pp. 19301–19325, 2018. doi: 10.1007/s11042-017-5317-2.

[19] Y. Liu, Y. Li, X. Yi, Z. Hu, H. Zhang and Y. Liu, "Micro-expression recognition model based on TV-L1 optical flow method and improved shufflenet," *Sci. Rep.*, vol. 12, no. 1, 2022, Art. no. 17522. doi: 10.1038/s41598-022-21738-8.

[20] S. -T. Liong, J. See, R. C. -W. Phan, K. Wong, and S. -W. Tan, "Hybrid facial regions extraction for micro-expression recognition system," *J. Signal Process. Syst.*, vol. 90, pp. 601–617, 2018. doi: 10.1007/s11265-017-1276-0.

[21] R. Ni, B. Yang, X. Zhou, S. Song, and X. Liu, "Diverse local facial behaviors learning from enhanced expression flow for microexpression recognition," *Knowl. Based Syst.*, vol. 275, 2023, Art. no. 110729. doi: 10.1016/j.knosys.2023.110729.

[22] X. Li, J. Yu, and S. Zhan, "Spontaneous facial micro-expression detection based on deep learning," in *IEEE. Int. Conf. Sig. Pro.*, Chengdu, China, Nov. 6–10, 2016, pp. 1130–1134.

[23] R. Chaudhry, A. Ravichandran, G. Hager, and R. Vidal, "Histograms of oriented optical flow and binet-cauchy kernels on nonlinear dynamical systems for the recognition of human actions," in *IEEE. Conf. Comp. Visi. Patt. Recog.*, Miami, FL, USA, Jun. 20–25, 2009, pp. 1932–1939.

[24] L. Zhou, Q. Mao, X. Huang, F. Zhang, and Z. Zhang, "Feature refinement: An expression-specific feature learning and fusion method for micro-expression recognition," *Pattern Recognit.*, vol. 122, 2022, Art. no. 108275. doi: 10.1016/j.patcog.2021.108275.

[25] S. -T. Liong, Y. S. Gan, J. See, H. -Q. Khor, and Y. -C. Huang, "Shallow triple stream three-dimensional CNN (STSTNet) for micro-expression recognition," in *IEEE Int. Conf. Auto.Fac. Recog.*, Lille, France, May 14–18, 2019, pp. 1–5.

[26] J. Li, T. Wang, and S. -J. Wang, "Facial micro-expression recognition based on deep local-holistic network," *Appl. Sci.*, vol. 12, no. 9, 2022, Art. no. 4643. doi: 10.3390/app12094643.

[27] N. Van Quang, J. Chun, and T. Tokuyama, "Capsulenet for micro-expression recognition," in *IEEE. Int. Conf. Auto.Fac. Recog.*, Lille, France, May 14–18, 2019, pp. 1–7.

[28] Z. Xia, X. Hong, X. Gao, X. Feng, and G. Zhao, "Spatiotemporal recurrent convolutional networks for recognizing spontaneous micro-expressions," *IEEE Trans. Multimedia*, vol. 22, no. 3, pp. 626–640, 2019. doi: 10.1109/TMM.2019.2931351.

[29] Y. Su, J. Zhang, J. Liu, and G. Zhai, "Key facial components guided micro-expression recognition based on first & second-order motion," in *IEEE. Int. Conf. Multi. Expo.*, Jul. 5–9, 2021, pp. 1–6.

[30] V. R. Gajjala, S. P. T. Reddy, S. Mukherjee, and S. R. Dubey, "MERANet: Facial micro-expression recognition using 3D residual attention network," in *Proc. Ind. Conf. Comp. Vsion. Grap. Imag. pro.*, Jodhpur, India, Dec. 19–22, 2021, pp. 1–10.

[31] H. Zhou, S. Huang, J. Li, and S. -J. Wang, "Dual-ATME: Dual-branch attention network for micro-expression recognition," *Entropy*, vol. 25, no. 3, 2023, Art. no. 460. doi: 10.3390/e25030460.

[32] H. -Y. Wu, M. Rubinstein, E. Shih, J. Guttag, F. Durand and W. Freeman, "Eulerian video magnification for revealing subtle changes in the world," *ACM Trans. Graph. (TOG)*, vol. 31, no. 4, pp. 1–8, 2012. doi: 10.1145/2185520.2185561.

[33] J. Liu, K. Li, B. Song, and L. Zhao, "A multi-stream convolutional neural network for micro-expression recognition using optical flow and EVM," 2020, *arXiv:2011.03756*.

[34] C. H. Yap, C. Kendrick, and M. H. Yap, "Samm long videos: A spontaneous facial micro- and macro-expressions dataset," in *IEEE. Int.Conf. Auto. Fac.Gest. Recog.*, Buenos Aires, Argentina, IEEE, 2020, pp. 771–776.

[35] T. Pfister, X. Li, G. Zhao, and M. Pietikäinen, "Recognising spontaneous facial micro-expressions," in *Int. Conf. Compu. Visi.*, Barcelona, Spain, Nov. 6–13, 2011, pp. 1449–1456.

[36] Q. Li, J. Yu, T. Kurihara, H. Zhang, and S. Zhan, "Deep convolutional neural network with optical flow for facial micro-expression recognition," *J. Circuits, Syst. Comput.*, no. 1, 2020, Art. no. 29. doi: 10.1142/S0218126620500061.

[37] W. Xu, H. Zheng, Z. Yang, and Y. Yang, "Micro-expression recognition base on optical flow features and improved MobileNetV2," *KSII Trans. Int. Inform. Syst.*, vol. 15, no. 6, pp. 1981–1995, 2021.

[38] C. Zach, T. Pock, and H. Bischof, "A duality based approach for realtime TV-L 1 optical flow," in *Pattern Recognition*, Springer, 2007, pp. 214–223. doi: 10.1007/978-3-540-74936-3_22.

[39] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," *Adv. Neural Inform. Process. Syst.*, vol. 27, pp. 568–576, 2014.

[40] S. Woo, J. Park, J. -Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Euro. Conf. Comput. Visio.*, Munich, Germany, Sep. 8–14, 2018, pp. 3–19.

[41] T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” in *Acm. sigk. Int. Conf. Knowl. Dis. Data. Min.*, San Francisco, CA, USA, Aug. 13–17, 2016, pp. 785–794.

[42] I. Tariq *et al.*, “Classification of pulsar signals using ensemble gradient boosting algorithms based on asymmetric under-sampling method,” *J. Instrum.*, vol. 17, no. 3, 2022, Art. no. P03020. doi: 10.1088/1748-0221/17/03/P03020.

[43] H. Xie, L. Lo, H. Shuai, and W. Cheng, “Au-assisted graph attention convolutional network for micro-expression recognition,” in *ACM. Int. Conf. Multi.*, Seattle, WA, USA, Oct. 12–16, 2020.

[44] J. Tang, L. Li, M. Tang, and J. Xie, “A novel micro-expression recognition algorithm using dual-stream combining optical flow and dynamic image convolutional neural networks,” *Signal, Image Video Process.*, vol. 17, no. 3, pp. 769–776, 2023. doi: 10.1007/s11760-022-02286-0.

[45] D. E. King, “Dlib-ml: A machine learning toolkit,” *J. Mach. Learn. Res.*, vol. 10, pp. 1755–1758, 2009.

[46] F. Chollet, “Xception: Deep learning with depthwise separable convolutions,” in *IEEE. Conf. Comp. Visi. Patt. Recog.*, Honolulu, HI, USA, Jul. 21–26, 2017.

[47] M. Abadi *et al.*, “Tensorflow: A system for large-scale machine learning,” in *Symp. Oper. Sys. Des. Imp.*, Savannah, GA, USA, Nov. 2–4, 2016, pp. 265–283.

[48] G. LemaÃžtre, F. Nogueira, and C. K. Aridas, “Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning,” *J. Mach. Learn. Res.*, vol. 18, no. 17, pp. 1–5, 2017.

[49] F. Pedregosa *et al.*, “Scikit-learn: Machine learning in python,” *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.

[50] P. Gupta, “MERASTC: Micro-expression recognition using effective feature encodings and 2D convolutional neural network,” *IEEE Trans. Affect. Comput.*, vol. 14, no. 2, pp. 1431–1441, 2021. doi: 10.1109/TAFFC.2021.3061967.

[51] Y. Li, X. Huang, and G. Zhao, “Joint local and global information learning with single apex frame detection for micro-expression recognition,” *IEEE Trans. Image Process.*, vol. 30, pp. 249–263, 2021. doi: 10.1109/TIP.2020.3035042.

[52] L. Zeng, Y. Wang, C. Zhu, W. Jiang, and J. Li, “Micro-expression recognition method combining dual-stream convolution and capsule network,” in *Int. Conf. Mach. Learn. Cyber. Sec.*, Yanuca Island, Fiji, Dec. 2–4, 2022, pp. 479–494.