Jordanian Journal of Computers and Information Technology (JJCIT), Vol. 11, No. 01, March 2025.

ENHANCING MICRO-EXPRESSION RECOGNITION: A NOVEL APPROACH WITH HYBRID ATTENTION-3DNET

Budhi Irawan[1,2], Rinaldi Munir[2], Nugraha Priya Utama[2] and Ayu Purwarianti[2]

(Received: 5-Sep.-2024, Revised: 11-Nov.-2024, Accepted: 12-Nov.-2024)

ABSTRACT

This paper proposes a unique pipeline for micro-expression recognition using a Dual-path 3D Convolutional Neural Network enhanced with Hybrid Attention and Squeeze-and-Excitation Blocks. The three main goals of the pipeline are to (1) Optimize the extraction of spatial-temporal features using advanced neural network architectures, (2) Enhance data representation by implementing targeted image augmentation and balanced class distribution and (3) Enhance feature fusion using state-of-the-art network techniques. Comprehensive experiments were conducted on four benchmark datasets: CAS(ME)^2, SMIC, SAMM and CASME II. The Hybrid Attention-3DNet model demonstrated superior recognition accuracy of 93.95% for CAS(ME)^2, 93.42% for SMIC, 93.61% for SAMM and 93.79% for CASME II, surpassing the state-of-the-art methods across these datasets. These outcomes demonstrate the efficacy and robustness of the proposed pipeline, underscoring its potential for a range of micro-expression recognition uses.

KEYWORDS

Micro-expression recognition, 3D convolutional dual-path network, Hybrid attention, Squeeze-and-excitation blocks, Deep learning.

1. INTRODUCTION

Micro-expressions are quick, uncontrollable facial movements that show true feelings that a person may try to hide. Even skilled observers may find it challenging to identify these expressions, since they are brief, frequently lasting less than 0.5 seconds [1]. Identifying micro-expressions has several uses, especially in security, psychology and medicine, where it is essential to comprehend genuine emotions [2].

New developments in deep learning have made it possible to create complex models to identify these nuanced expressions. Nevertheless, fundamental difficulties persist, including a lack of data and a notable disparity in micro-expression classes [3]. Furthermore, conventional Convolutional Neural Networks (CNNs) frequently demand data and require assistance with overfitting in situations where data is limited [4].

This study suggests a unique architecture that combines Hybrid Attention and Squeeze-and-Excitation Blocks into a Dual-path 3D Convolutional Neural Network (3DCNN) to improve micro-expression identification. Tested on benchmark datasets, including CAS(ME)^2,[5] SMIC [6], SAMM [7] and CASME II [8], the suggested model outperformed state-of-the-art techniques in terms of accuracy, indicating its potential for practical use in emotion recognition [9].

2. RELATED WORK

Recent developments in micro-expression recognition have sparked the creation of creative techniques to increase precision. Combining CNNs with other methods is one such strategy. A technique that combines Swin Transformer and ConvNeXt is presented in [10] and is based on a Dual-branch Spatiotemporal Convolutional Network (STCN). This method uses both CNN and Transformers to address issues, including the preservation of facial spatial structure and the localization of micro-expression actions. According to tests on the CASME and SMIC datasets, the STCN network improves micro-expression identification accuracy.

The Divided-block Multi-scale Convolution Network (DBMNet) is a unique multi-scale convolutional

1. B. Irawan is with Telkom University, Indonesia. Email: budhiirawan@telkomuniversity.ac.id
2. B. Irawan, R. Munir, N. P. Utama and A. Purwarianti are with Bandung Institute of Tecnology, Indonesia. Emails: budhiirawan@telkomuniversity.ac.id, rinaldi@informatika.org, utama@informatika.org and ayu@informatika.org

network proposed in another study [11]. This network is intended to learn from four different optical flow feature images produced between the micro-expression samples' onset and apex frames. With the use of the Divided-block Multi-scale Convolution Module (DBMCM), the network can efficiently capture more intricate and useful multi-scale properties of micro-expressions.

A deep-learning technique known as the Spatiotemporal Capsule Network (STCP-Net) was recently presented in [12]. This method aims to increase recognition accuracy while decreasing recognition time. The four main parts of STCP-Net are a jitter removal module, a differential feature-extraction module, a spatiotemporal capsule module and a fully connected layer.

[13] presents the Parallel Dual-branch Attention-based Spatio-temporal Fusion Network (PASTFNet). This method is influenced by the combined architecture of Long-Short-Term Memory (LSTM) and CNN for temporal modeling. The paper suggests encoding sequential frame features using an attention-based multi-scale feature-fusion network (AMFNet). The network gathers more expressive face-detail features for micro-expression recognition through multi-scale feature fusion and integrated attention.

A two-layer feature-encoding technique is suggested in [14] to depict interactions across different regions of the feature map, along with a novel multi-frame technique intended to capture subtle motion patterns. The paper also presents an Action Unit Graph Convolutional Network (AU GCN). It uses a transformer encoder, an adjacency matrix and an AU-detection module to adjust to test data.

A Triple-branch Attention Fusion Network (Triple-ATFME) is presented in [15] for micro-expression recognition. With the help of a Triple-branch ShuffleNet module, an adaptive channel attention module and pre-processing, this approach enables the model to extract multi-view features using a multi-path architecture. The framework uses optical-flow approaches to capture various optical-flow information by extracting optical-flow features from the facial region's cropped start and peak frames. The Triple-ATFME network processes these features to find hidden features. Channel features are adjusted *via* a Channel Fusion Attention Module (CFAM) to improve multi-view feature integration and lessen the model's emphasis on local information during feature fusion.

Lastly, the suggested framework in this research introduces a unique method for improving the accuracy of micro-expression identification through spatio-temporal deep learning, data augmentation and class balancing [16]. Rotation, contrast adjustment and SMOTE are examples of sophisticated pre-processing techniques that the framework uses to effectively handle data restrictions and class imbalances, enhancing model generalization and lowering overfitting. Based on the results, this method set a new standard for micro-expression analysis and created a more dependable model for emotion-detection applications. It also made notable gains, especially in accuracy and F1-scores across many datasets. In addition, this report serves as a baseline for the research.

3. PROPOSED METHOD

This study adopts a systematic approach by developing a pipeline to classify input video datasets of spontaneous micro-expressions. The datasets are categorized into three emotional classes: angry, happy and disgusted for the CAS(ME)² dataset and positive, negative and surprise for the SMIC, SAMM and CASME II datasets. The study is structured around four experimental scenarios, each tailored to the input from these four datasets.

The developed pipeline consists of several stages, including data preparation, pre-processing, classification and performance measurement, as illustrated in Figure 1. The datasets are organized based on their respective emotional classes in the data-preparation stage. Specifically, the CAS(ME)² dataset is divided into three classes: angry, happy and disgusted, while the SMIC, SAMM and CASME II datasets are categorized into positive, negative and surprise classes.

During pre-processing, the video clips from the datasets are converted into a series of image frames, followed by face detection and the identification of 68 facial landmarks. The areas around the eyes and mouth are masked and the face is cropped. The images are then resized to 128x128 pixels and converted into grayscale. Additionally, data augmentation is performed by adjusting orientation, contrast and brightness and class balancing is achieved using the class-weight method. The pre-processed facial images are subsequently divided into upper and lower face regions.

This study proposes a model for the classification task that utilizes a dual-path 3D convolutional neural

network incorporating hybrid attention and Squeeze-and-Excitation blocks. Finally, the emotion classification is evaluated using accuracy, F1-score and error rate, ensuring a comprehensive assessment of the model's performance.

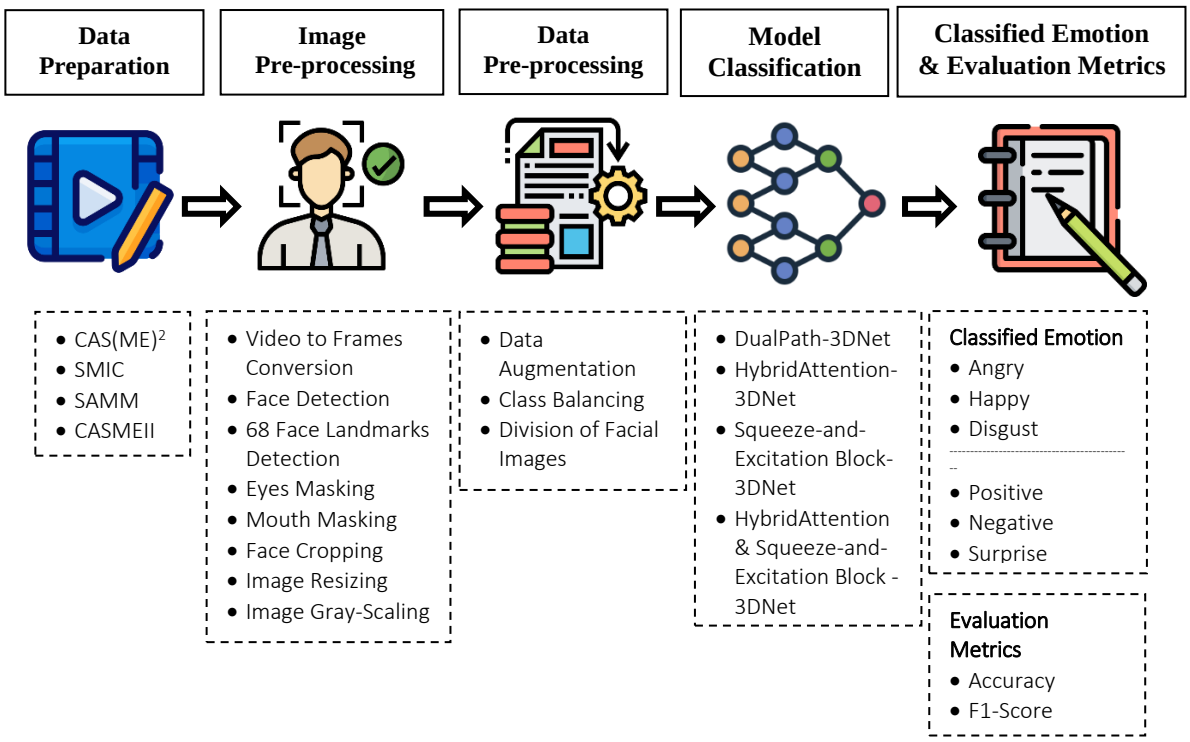

Figure 1. The proposed design of a micro-expression recognition pipeline.

3.1 Pre-processing Stages

The process begins with image pre-processing to identify spontaneous micro-expressions from extended sequences of facial videos. Initially, the emotional classes of each dataset are categorized. The CAS(ME)² dataset includes categories for emotions, such as anger, happiness and disgust, while the SMIC, SAMM and CASME II datasets cover positive, negative and surprise emotions. In the next step, video clips from these datasets are converted into sequential image frames. The original resolutions of these frames vary by dataset: 640x480 pixels for CAS(ME)² and SMIC, 960x560 pixels for SAMM and 280x340 pixels for CASME II, as depicted in Figure 2.

The provided diagram demonstrates the sequential steps of image and data pre-processing, beginning with raw video input and concluding with facial-image sequences split into upper and lower sections, each resized to 128x64 pixels. The main goal of pre-processing is to optimize raw video data for practical use in micro-expression recognition. This process transforms video data into clean, structured image sequences that highlight essential facial features, reduce noise and maintain a balanced distribution of classes. It begins by converting the raw video into individual frames, with each dataset's frames retaining specific resolutions: 640x480 pixels for CAS(ME)² and SMIC, 960x560 pixels for SAMM and 280x340 pixels for CASME II. Face detection is applied to each frame during pre-processing, followed by identifying 68 facial landmarks. These landmarks outline critical facial areas, such as the eyes, mouth and nose, guiding the masking and segmentation steps necessary for accurate micro-expression recognition.

Once the landmarks are detected, the eye and mouth areas are masked to focus on the most expressive facial regions, which aids the model in capturing subtle movements associated with micro-expressions. The face is then cropped, resized to 128x128 pixels and converted into grayscale to simplify color-data without sacrificing critical information, thereby speeding up analysis while preserving accuracy. Data augmentation techniques, including rotation, cropping and contrast adjustments, are also applied to enhance data variability. This step increases the diversity of the dataset, helping the model to generalize across various angles and lighting conditions.

Class balancing is implemented to ensure a balanced representation across classes by assigning weights to each class, reducing bias towards majority classes and improving the model's ability to recognize minority classes accurately. Finally, the pre-processed facial images are divided into upper and lower sections, each resized to 128x64 pixels and organized into frame sequences ready for classification. This segmentation enables the model to analyze distinct facial areas independently, enhancing the detection of subtle changes in the eyes and mouth regions that play a crucial role in micro-expression recognition.

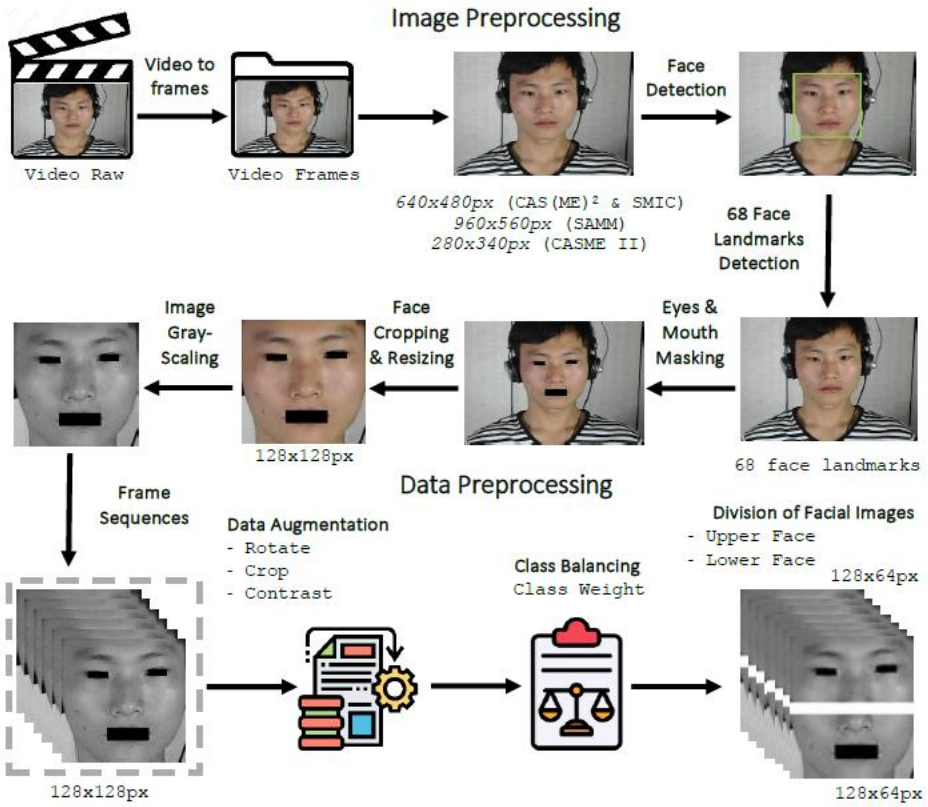

Figure 2. Pre-processing stages.

3.2 Data Optimization

The quality of a dataset is crucial in ascertaining the accuracy and efficiency of machine-learning models in data processing. Imbalanced or unrepresentative data may result in biased models and suboptimal performance. Consequently, diverse methodologies enhance data, enabling models to learn more efficiently and generate more precise predictions. Data augmentation and class balance are two essential methodologies employed in this process, which will be examined in more detail in the subsequent sections. Data-augmentation techniques are essential in machine learning and image processing, particularly for improving the quality and quantity of training datasets. These strategies entail methodically modifying existing photos to create new variants and augmenting the dataset without further data collection. Data augmentation is indispensable in micro-expression recognition, where extensive datasets are vital for enhancing model accuracy.

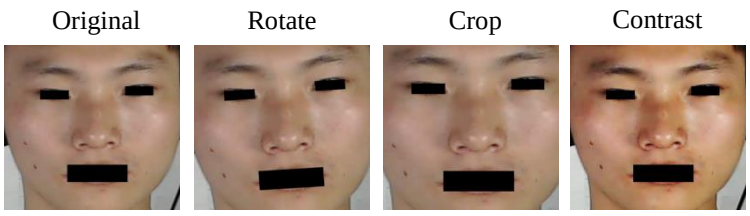

Figure 3. Data augmentation.

In this context, image frames can undergo various transformations, including rotation, cropping and modifications to brightness and contrast, as illustrated in Figure 3. The transformations generate varied representations of the original images, enhancing the model's ability to recognize and adapt to diverse

patterns and variations in the data. Data augmentation enhances model generalization by increasing the diversity of training samples, resulting in improved performance across various scenarios.

Class balancing is a technique employed to equalize the distribution of samples or observations across various classes within a dataset. In classification tasks, a class denotes the specific label or category that the model aims to predict. An imbalance occurs when there is a significant disparity in the number of samples between classes. The imbalance in datasets presents considerable challenges in machine learning, as models trained on such uneven data frequently produce biased predictions, favoring the majority classes while overlooking minority groups. This bias may hinder the model's capacity to identify and classify instances from the minority classes accurately. This study implements a method of class balancing known as the class weight method.

The class weight method is instrumental in detecting micro-expressions, particularly by addressing distribution imbalances in the dataset. In micro-expression recognition, class imbalance is a common issue; some classes have more samples than others. By assigning higher class weights to the less common classes, the model is encouraged to learn from the minority classes more effectively. This approach is beneficial in mitigating class imbalance and enhancing the model's ability to recognize minority classes. Furthermore, rare classes, such as specific facial expressions that occur infrequently, typically exhibit lower accuracy due to the limited number of samples [17]. Assigning class weights offers greater motivation for the model to accurately recognize these classes, thereby improving its performance on minority classes.

In addition, applying class weights is anticipated to minimize the risk of overfitting to the majority class. Assigning weights to each class prevents the model from overly focusing on the majority class and promotes a more balanced learning process across all classes. The equation used for calculating class weights in cases of class imbalance often involves comparing the sizes of the classes or employing simple proportional methods. The general formula for computing class weights is given by Equation (1):

$$W_k = \frac{N}{n_k} \tag{1}$$

where W_k is the class weight for class k, N is the total number of samples in the training data and n_k is the number of samples in class *k*.

This equation implies that the minority class will have a higher class weight than the majority class. This inverse relationship between the sample proportion within the class and the assigned class weight gives the minority class more importance in the learning process, aiding in overcoming class imbalance.

3.3 Division of Facial Images

The Division of Facial Images is crucial in enhancing micro-expression recognition by focusing on specific regions of facial-muscle activity. In micro-expressions, Action Units (AU) are located in distinct facial regions where subtle muscle movements occur, often concentrated around the eyes, eyebrows and mouth. These areas contain a dense, exemplary muscle network that enables intricate facial movements. For example, muscle contractions around the eyes can form wrinkles, while those around the mouth can alter lip shape. This division aims to develop a more precise and focused approach to facial micro-expression recognition by acknowledging the significance of AU locations.

In the context of a dual-stream input classification model for facial-expression recognition, dividing the facial image into upper and lower sections enables the model to concentrate on the specific distribution of AUs across the face. This division facilitates the separate processing of the upper and lower face regions, aligning with their distinct roles in expressing emotions. The original image, sized at 128x128 pixels, is split into two parts, each measuring 128x64 pixels. According to Ekman's research on facial regions and emotional expression, the upper face, which includes the eye and eyebrow areas, is predominantly associated with emotions such as positivity and surprise. Conversely, the lower face, encompassing the nose, mouth and cheeks, often conveys subtler or more complex emotional nuances, particularly negative emotions.

The process of dividing facial images in this way enables the model to capture spatio-temporal features related to micro-expression AUs more effectively. By processing these sections separately, the dual-stream classification model can focus on the distinct emotional signals conveyed by each part of the

face. This separation allows the model to detect fine-grained emotional expressions, improving accuracy in recognizing micro-expressions across different facial regions. The illustration in Figure 4 demonstrates the process of dividing the facial image into upper and lower sections, each resized to 128x64 pixels, enhancing the model's capacity to analyze and classify these regions independently. Thus, the Division of Facial Images facilitates the targeted analysis of key facial regions and contributes to a more nuanced understanding of emotion-expression dynamics, which is essential for effective micro-expression recognition.

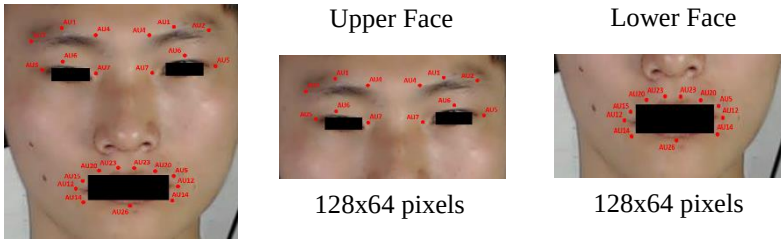

Figure 4. Division of facial images.

3.4 The 3D Convolutional Dual Path Network Model

This study utilizes a 3D convolutional dual-path network model for micro-expression recognition, which enhances attention to spatial-temporal feature weights. The model incorporates hybrid attention and squeeze-and-excitation blocks to recognize spontaneous micro-expressions. This dual-path model is an extension of the single-path model [18], which primarily focuses on general facial features without addressing the detailed features of the upper and lower facial regions. The proposed research pipeline includes four model designs: Dual Path-3DNet, Hybrid Attention-3DNet, SE Block-3DNet and Hybrid Attention-3DSENet.

3.4.1 Dual Path-3DNet

The purpose of proposing the Dual Path-3DNet model is to enhance the recognition of micro-expressions by utilizing a dual-stream approach that separately processes different facial regions, thereby capturing more detailed spatial-temporal features relevant to each region. Using two distinct input paths, this model can independently analyze the upper and lower sections of the face derived from the pre-processing steps applied to the four datasets used in this study. This dual-path strategy allows the model to focus on region-specific characteristics in each section, optimizing the detection of subtle emotional cues that may be localized to particular facial areas.

In the Dual Path-3DNet model's design, each input path is structured to process sequences of pre-processed image frames and the segmented upper-face and lower-face regions. Each path includes layers, including 3D convolution with ReLU activation, 3D max pooling, flatten, dense with ReLU and dropout layers. These layers enable effective feature extraction and reduce overfitting by discarding non-essential data points. The outputs of the two paths are then merged into a single pathway through a concatenate layer, which integrates the distinct information extracted from both facial regions.

Following this merging, the combined pathway passes through additional dense layers with ReLU activation, dropout and finally, a softmax layer for classification. This final pathway enables the model to learn complex interrelations between features from both facial regions, allowing for more accurate and nuanced emotion detection. The architecture of the Dual Path-3DNet model is illustrated in Figure 5, where each layer and process flow is visually represented to demonstrate the interaction between the dual pathways. By adopting this dual-path approach, the model is better equipped to recognize micro-expressions by simultaneously analyzing diverse facial features across regions. This ultimately contributes to higher accuracy in emotional-recognition tasks.

3.4.2 Hybrid Attention-3DNet

The proposed design of the Hybrid Attention-3DNet model shares the same basic architecture as the Dual Path-3DNet model. The critical difference in this architecture is the addition of a hybrid-attention layer (encompassing both Spatial and Temporal Attention) placed after the 3D max pooling layer. This

is followed by flatten, dense + ReLU and dropout layers, which are then merged into a single path using a concatenate layer. After this merging process, the subsequent path consists of dense layers with ReLU activation, dropout and a softmax layer. The architecture of the Hybrid Attention-3DNet model is illustrated in Figure 6.

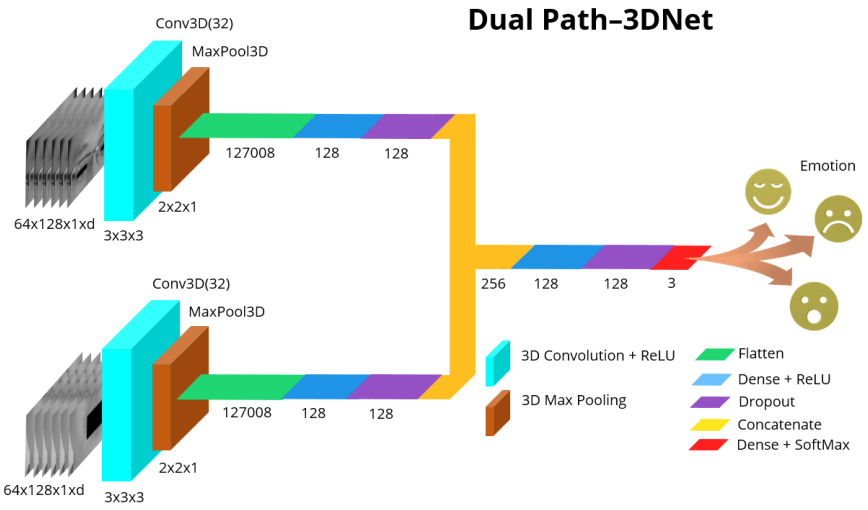

Figure 5. The architecture of the dual path-3DNet model.

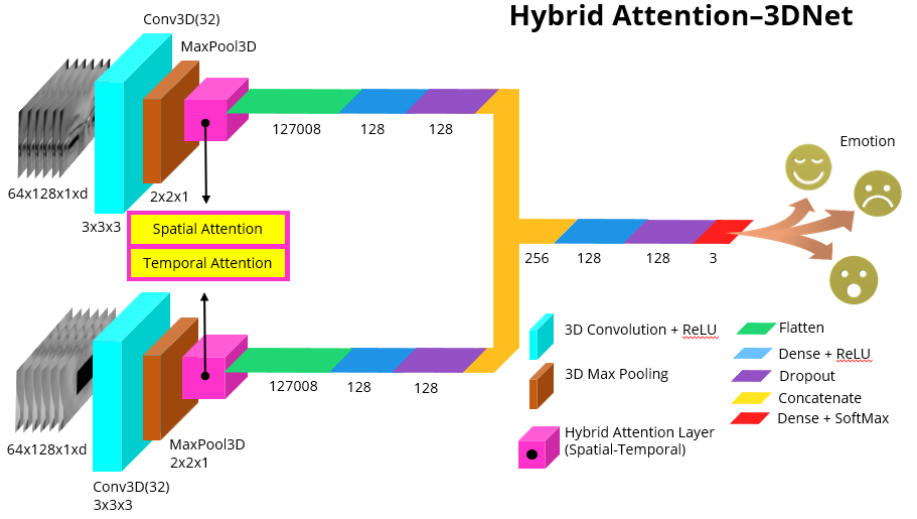

Figure 6. The architecture of the hybrid attention-3DNet model.

This model incorporates Hybrid Attention, composed of Spatial and Temporal Attention, to enhance its ability to recognize micro-expressions. Hybrid Attention is applied following the 3D max pooling layer, where each attention mechanism, spatial and temporal, contributes to the improved representation of spatial and temporal features.

Spatial attention is designed to identify critical spatial features within facial regions. This mechanism increases the weights for crucial areas, such as those around the eyes and mouth, using an attention map to determine the distribution of activations across spatial features. By selectively amplifying significant features, the model is better equipped to identify specific patterns linked to micro-expressions that could be challenging to detect without spatial attention.

Temporal attention focuses on capturing sequential changes in facial expressions, allowing the model to detect subtle variations that occur from frame to frame over short durations in micro-expression videos. This layer assigns weights based on temporal dynamics, enabling the model to recognize small changes in facial expressions that might be overlooked by traditional methods that do not incorporate temporal information.

The integration process and benefits from Hybrid Attention are incorporated after the 3D max pooling

layer and positioned before the dual-path 3DCNN concatenation stage. This layer processes inputs from the refined feature maps, ensuring that spatial and temporal attention mechanisms are employed before the final classification step. Through this combination, the model can prioritize essential features in both spatial and temporal dimensions, thus enhancing responsiveness to rapid, subtle variations in expressions, leading to more accurate micro-expression detection in sequential video data.

3.4.3 Squeeze-and-Excitation Block-3DNet

Incorporating the Squeeze-and-Excitation Block-3DNet model enhances feature selection by emphasizing critical spatial-temporal information within the data, which is essential for accurate micro-expression recognition. While the SE Block-3DNet model shares the basic structure with the Dual Path-3DNet model, it introduces a Squeeze-and-Excitation block after the 3D max pooling layer. This additional layer selectively emphasizes significant features by applying global average pooling (Global AP) to squeeze the spatial dimensions, followed by fully connected layers with ReLU and sigmoid activation to recalibrate the feature-map channels.

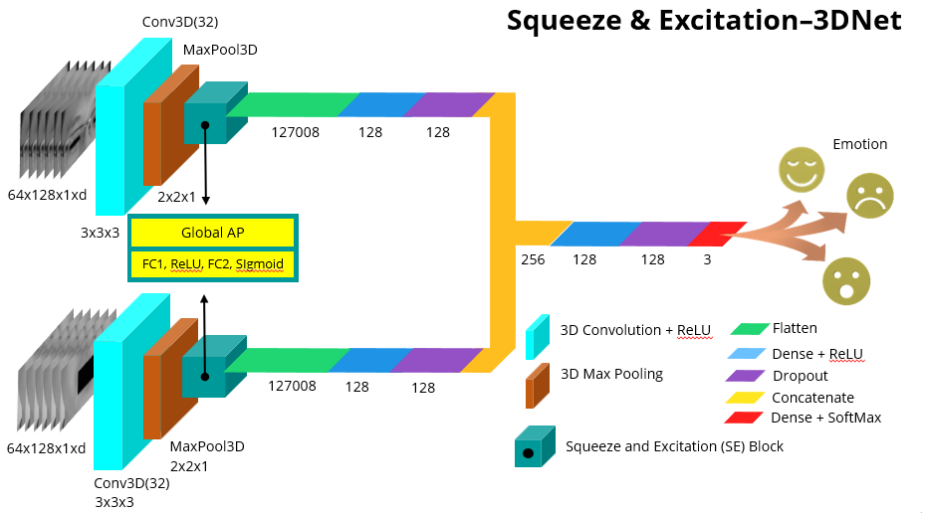

Figure 7. The architecture of the squeeze & exitation-3DNet model.

After recalibration, the Squeeze-and-Excitation Block's output undergoes the same layers as the Dual Path-3DNet model, including flatten, dense with ReLU and dropout layers. The pathways from each input are then merged using a concatenate layer. Further dense layers with ReLU activation, dropout and a softmax layer for final classification follow this merging. The Squeeze-and-Excitation Block's selective attention mechanism helps the model focus on essential micro-expression cues, optimizing the feature representation for each facial region.

As shown in Figure 7, the SE Block-3DNet model's architecture incorporates this additional Squeeze-and-Excitation Block, which enhances the model's ability to prioritize essential features, leading to improved performance in emotion-recognition tasks. This approach leverages spatial recalibration and dual-path processing to capture fine-grained details, making it a powerful method for precise micro-expression analysis.

3.4.4 Hybrid Attention-3DSENet

Introducing the Hybrid Attention Squeeze-and-Excitation Block-3DNet model aims to enhance the model's ability to capture both spatial and temporal features essential for recognizing micro-expressions. This model builds on the fundamental structure of the Dual Path-3DNet, but incorporates a hybrid attention layer that combines both spatial and temporal attention mechanisms. Placed after the 3D max pooling layer, this hybrid-attention layer selectively focuses on important spatial locations and temporal sequences, optimizing feature extraction for subtle micro-expressions.

After the attention layer, the model continues with flatten, dense + ReLU and dropout layers, which are subsequently merged using a concatenate layer to integrate features from both input paths. Following this merge, a Squeeze-and-Excitation Block layer is added, providing further refinement by recalibrating

channel importance and is then followed by dense layers with ReLU activation, dropout and finally, a softmax layer for classification.

As illustrated in Figure 8, the Hybrid Attention-3DSENet architecture effectively combines spatial and temporal attention with channel recalibration through the Squeeze-and-Excitation Block. This integration allows the model to prioritize essential micro-expression cues across both dimensions, producing more precise and robust emotion recognition. This hybrid-attention mechanism and dual-path processing make the model particularly effective for detailed micro-expression analysis.

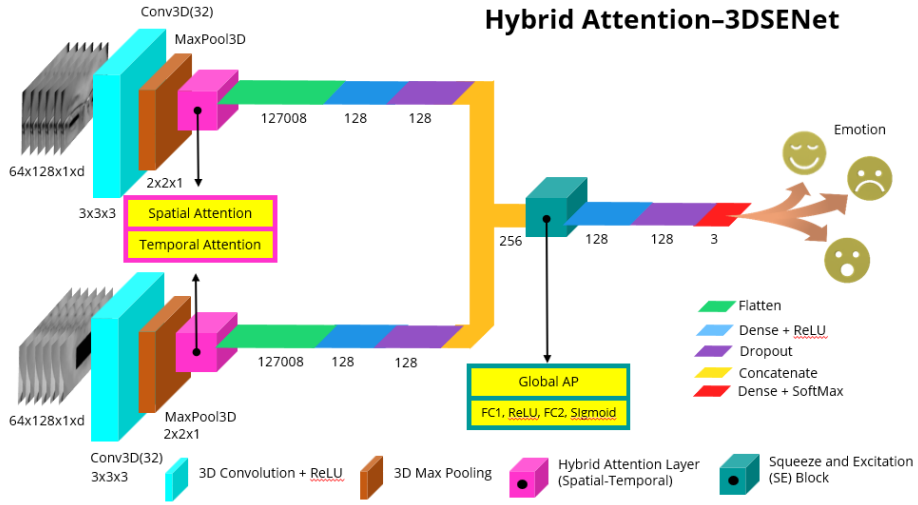

Figure 8. The architecture of the squeeze-and- exitation-3DNet model.

3.5 Hybrid Attention and Squeeze-and-Excitation Block

The attention mechanism is a technique in artificial neural networks that enables the model to focus on specific input parts when making predictions. This technique is beneficial in tasks, such as pattern recognition and object detection in vision systems. The mechanism assigns different weights to input elements, allowing the model to emphasize more relevant information while disregarding less essential details. The primary benefit of the attention mechanism is its ability to improve model performance by capturing more complex contexts and reducing computational load by concentrating resources on the most crucial information. Consequently, the model becomes more efficient and accurate in processing large and heterogeneous datasets.

One variant of the attention mechanism is hybrid attention. Hybrid attention is an approach that combines spatial and temporal attention in a 3D convolutional network model to enhance performance in tasks, such as image or video classification [19]. The primary function of the Hybrid Attention layer is to enable the model to capture important information spatially and temporally from the input data. The Hybrid Attention layer is illustrated in Figure 9.

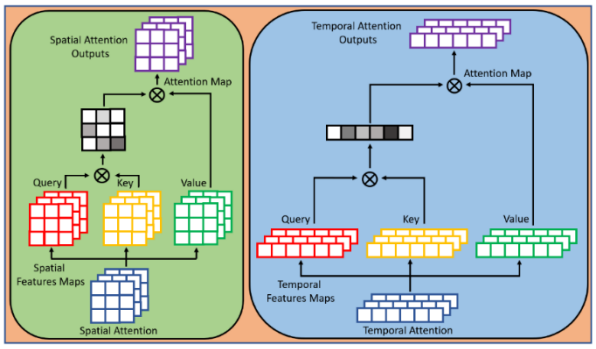

Figure 9. Hybrid attention.

The Squeeze-and-Excitation Block is widely adopted in convolutional neural networks to enhance the model's ability to extract significant image features. The Squeeze-and-Excitation Block layer functions

by assigning higher weights to important features while reducing the weights of less relevant ones. The Squeeze-and-Excitation Block operates through two main stages. The first stage, the "squeeze" stage, involves a global average pooling operation to generate descriptors or feature vectors representing the aggregate information from all feature channels. This stage reduces dimensionality and complexity, producing a more compact, yet informative, representation.

The next stage, the excitation stage, uses the feature vector generated from the squeeze stage and passes it through a series of fully connected layers, including a ReLU activation layer and a sigmoid layer. The ReLU layer enhances the representational capacity and flexibility in learning feature relationships. In contrast, the sigmoid layer produces weights or scalars that indicate the importance of each feature channel.

The Squeeze-and-Excitation Block shows the process flow from input to output within the Squeeze-and-Excitation Block layer. First, the input with dimensions HxWxC is fed into the global average pooling stage to generate feature descriptors. These descriptors then pass through two fully connected layers with ReLU and sigmoid activations, producing scalars representing each feature channel's significance. These scalars are subsequently used to scale the original features through a residual operation, resulting in an adjusted output.

By passing features through the Squeeze-and-Excitation Block, the model can adaptively select and focus attention on the most relevant and essential features within the image while disregarding less informative ones. This process helps the model obtain more robust and discriminative representations from the input data, ultimately improving performance in classification tasks. The Squeeze-and-Excitation Block is typically placed after the convolutional layer in a convolutional neural network architecture. This enables the model to effectively capture spatial-temporal features from images or sequences and apply more significant attention to the most critical features using the Squeeze-and-Excitation Block.

4. EXPERIMENT SETUP

Defining the experiments' scope within the context of developing a micro-expression recognition classification model is crucial for ensuring the model's effectiveness and generalizability. This is essential to guarantee that the experiments conducted are relevant to real-world conditions and can provide meaningful solutions to practical problems. Furthermore, the experiments' scope encompasses various scenarios and micro-expression variations, ensuring that the model can manage emotional differences' complexity. The scope also facilitates hyper-parameter optimization, enabling fine-tuning to achieve the most effective configuration.

Hyper-parameter tuning is conducted to identify the optimal combination of parameters that maximizes model performance, especially in micro-expression recognition. This involves experimenting with different parameters to find the configuration that delivers the most effective results. In this study, parameters such as batch sizes of 80 or 100 are selected to ensure comprehensive data processing, reducing the likelihood of missing critical patterns and helping prevent overfitting. Using 200 or 250 epochs allows for early monitoring of model performance, which helps avoid unnecessary overtraining. Data is divided into 80% for training, 10% for testing and 10% for validation in order to guarantee a thorough model evaluation.

Implementing the Adaptive Moment Estimation (ADAM) optimizer is essential for adjusting the learning rate in the intricate 3D convolutional neural network, resulting in faster and more stable convergence. The default learning rate of 0.001 balances stability and convergence speed. The Categorical Cross-Entropy loss function effectively manages multi-category classification tasks, ensuring precise facial micro-expression identification. This strategic combination of parameters and methods enhances the model's generalizability and improves its effectiveness in recognizing micro-expressions.

An experimental scenario is a series of plans and steps that are iteratively and alternately conducted to achieve the desired outcomes. This study's four experimental scenarios correspond to the classification models developed: Dual path-3DNet, Hybrid Attention-3DNet, SEBlock-3DNet and Hybrid Attention-3DSENet. Each scenario employs one dataset and applies image pre-processing, data pre-processing and variations in batch size and epoch. Each dataset undergoes 16 experiments, resulting in 64

experiments conducted across the four datasets in this study. A detailed explanation of each experimental scenario is provided in Table 1.

Table 1. Experimental scenarios.

Scenario	Dataset	Pre-processing	Augmentation	Class Weight	Batch Size	Epoch	Hybrid Att.	SE-Block
1	CAS(ME)^2	✓	✓	✓	80/100	200/250	×	×
	SMIC	✓	✓	✓	80/100	200/250	×	×
	SAMM	✓	✓	✓	80/100	200/250	×	×
	CASME II	✓	✓	✓	80/100	200/250	×	×
2	CAS(ME)^2	✓	✓	✓	80/100	200/250	✓	×
	SMIC	✓	✓	✓	80/100	200/250	✓	×
	SAMM	✓	✓	✓	80/100	200/250	✓	×
	CASME II	✓	✓	✓	80/100	200/250	✓	×
3	CAS(ME)^2	✓	✓	✓	80/100	200/250	×	✓
	SMIC	✓	✓	✓	80/100	200/250	×	✓
	SAMM	✓	✓	✓	80/100	200/250	×	✓
	CASME II	✓	✓	✓	80/100	200/250	×	✓
4	CAS(ME)^2	✓	✓	✓	80/100	200/250	✓	✓
	SMIC	✓	✓	✓	80/100	200/250	✓	✓
	SAMM	✓	✓	✓	80/100	200/250	✓	✓
	CASME II	✓	✓	✓	80/100	200/250	✓	✓

5. EVALUATION METRICS

In classification evaluation, specific metrics are used to evaluate a model's ability to predict the target class of a dataset. Essential terms include True Positives, False Positives, True Negatives and False Negatives. A True Positive (TP) occurs when the model correctly predicts a positive instance, aligning with the actual class. In essence, TP represents the count of positive samples accurately identified. A False Positive (FP), on the other hand, happens when the model incorrectly predicts a negative sample as positive, reflecting the number of negative instances misclassified as positive. True Negative (TN) refers to instances where the model correctly classifies a sample as negative, matching the actual class and representing the accurate identification of negative instances. Finally, a False Negative (FN) occurs when the model incorrectly predicts a positive sample as negative, signifying the number of positive instances mistakenly identified as negative.

Evaluation metrics are specific measures to gauge a deep-learning model's performance. Choosing the right metric is essential, as it influences the assessment of the model's ability to complete the task and helps compare the efficiency of various models.
Accuracy measures how well a classification model identifies the correct classes across the entire dataset [20]. It is determined by dividing the sum of true positives and true negatives by the total number of samples.

$$Accuracy = \frac{True\ Positives + True\ Negatives}{\text{Total Samples}} \quad (2)$$

The F1-score is a metric used to evaluate a model by balancing precision and recall. Precision measures the accuracy of the model's positive predictions, while recall evaluates how well the model detects all true positive instances.

$$F1 - score = 2\ \text{x}\ \frac{Precission x Recall}{Precission + Recall} \quad (3)$$

Recall assesses a model's capability to detect all actual positive instances. It is determined by dividing

the number of true positives by the total of true positives and false negatives. A high recall indicates that the model effectively captures the majority of positive cases within the dataset.

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives} \quad (4)$$

Precision measures how accurately a model predicts positive outcomes when they are indeed positive. It is calculated by dividing the true positives by the total number of predicted positive instances, which includes both true positives and false positives. A high precision score shows that the model makes few false positive predictions.

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives} \quad (5)$$

The error rate quantifies how often a model misclassifies data points. It is determined by dividing the sum of false positives and false negatives by the total number of samples. A lower error rate reflects that the model makes few classification mistakes.

$$Error\ Rate = \frac{False\ Positives + False\ Negatives}{Total\ Samples} \quad (6)$$

6. RESULTS AND DISCUSSION

The MER-DACWB3DCNNST model (Single-path 3DCNN) serves as the baseline in this study [18], offering a straightforward, yet practical, approach for micro-expression recognition by leveraging spatio-temporal features from a single input stream of facial images to enhance accuracy in detecting subtle micro-expressions. This model achieves strong results with a relatively simple structure while minimizing computational complexity. However, its primary limitation lies in its constrained ability to capture the depth of complex micro-expression features, particularly in cases where expression variations are incredibly subtle and rapid. This limitation can reduce the model's performance, detecting nuanced and fleeting emotional changes. Nonetheless, the model demonstrates robust accuracy and F1 score results across several micro-expression datasets, as shown in Table 2.

Table 2. Comparison of accuracy and F1-score of the MER-DACWB3DCNNST model (single-path 3DCNN) across different datasets.

Dataset	Accuracy (%)	F1-score
CAS(ME)²	92.75	0.9271
SMIC	91.49	0.9032
SAMM	92.20	0.9218
CASME II	93.66	0.9361

To enhance model performance, this study incorporated three additional components for testing: the Dual-path 3D CNN model, Hybrid Attention and Squeeze-and-Excitation Blocks, each utilizing dual input streams from the upper and lower facial regions. An ablation study was conducted to assess the impact of each component on model performance, using accuracy and F1-score as the primary metrics across multiple datasets. This analysis provides insights into the contribution of each component compared to the single-path baseline model.

Evaluation metrics such as accuracy and F1-score are calculated in each experimental stage. The presented graphs include information from all experimental scenarios, covering the dataset used, the application of image pre-processing and data pre-processing, batch size and epoch selection, as well as the accuracy and F1-score values. The type of graph presented is a line graph, which displays the highest accuracy and F1-score values for each experiment based on the dataset used. To provide detailed insights into the accuracy and F1-score calculations for each experimental scenario, the graphs are accompanied by a complete table of the experimental results. From these graphs, conclusions and analyses related to the obtained results can be drawn.

Figure 10 presents the accuracy and F1-score graphs for micro-expression recognition using four datasets: CAS(ME)², SMIC, SAMM and CASME II, with the proposed models: Dual Path-3DNet, Hybrid Attention-3DNet, Squeeze-and-Excitation-3DNet and Hybrid Attention-3D Squeeze-and-Excitation Net. From the graphs, Hybrid Attention-3DNet (HA-3DNet) model consistently achieves the highest accuracy across all datasets, with scores of 93.95% for CAS(ME)², 93.42% for SMIC, 93.61% for SAMM and 93.79% for CASME II. F1-scores are similarly high across datasets, with 0.9395 for

CAS(ME)², 0.9330 for SMIC, 0.9113 for SAMM and 0.9203 for CASME II.

The Dual-path 3DCNN structure performed better than the baseline Single-path model, particularly in capturing spatial-temporal features within different facial regions. The Dual-path approach enhances the model's sensitivity to subtle spatial-temporal dynamics by enabling separate pathways for upper and lower face regions. Results indicate a consistent accuracy improvement across all datasets, suggesting that the Dual-path structure provides more robust feature extraction.

Incorporating Hybrid Attention into the Dual-path 3DCNN model, which combines spatial and temporal attention mechanisms, further enhances the model's performance. This component allows the model to prioritize critical facial features like eyes and mouth while adapting to temporal variations. Experiments indicate that adding Hybrid Attention significantly boosts both accuracy and F1-score, highlighting its effectiveness in refining feature relevance. This component has proven especially effective on datasets with subtle, rapid expressions, confirming its essential role in distinguishing fleeting emotions.

Including Squeeze-and-Excitation Blocks in the Dual-path 3DCNN model enables adaptive reweighting of feature channels, allowing the model to highlight the most relevant features while downplaying less significant elements. This mechanism helps reduce noise in the micro-expression recognition process, particularly for complex expressions involving subtle changes across various facial regions. Squeeze-and-Excitation Blocks have proven effective in enhancing classification precision, as reflected in increased F1-scores across all datasets. The contribution of the Squeeze-and-Excitation Blocks is especially evident in recognizing expressions that require high sensitivity to specific features, delivering consistent and stable results in micro-expression classification.

Integrating Dual-path 3DCNN, Hybrid Attention and Squeeze-and-Excitation Blocks, the whole combination model achieves high accuracy and F1-scores across all datasets, indicating that each component contributes meaningfully to the model's overall performance. This combination offers robust feature extraction, adaptive focus and refined feature weighting. However, as shown in the result graphs, the model using only Hybrid Attention on Dual-path 3DCNN outperforms the whole combination, suggesting that the contribution of each component in this combination does not necessarily yield a more significant performance boost than Hybrid Attention alone.

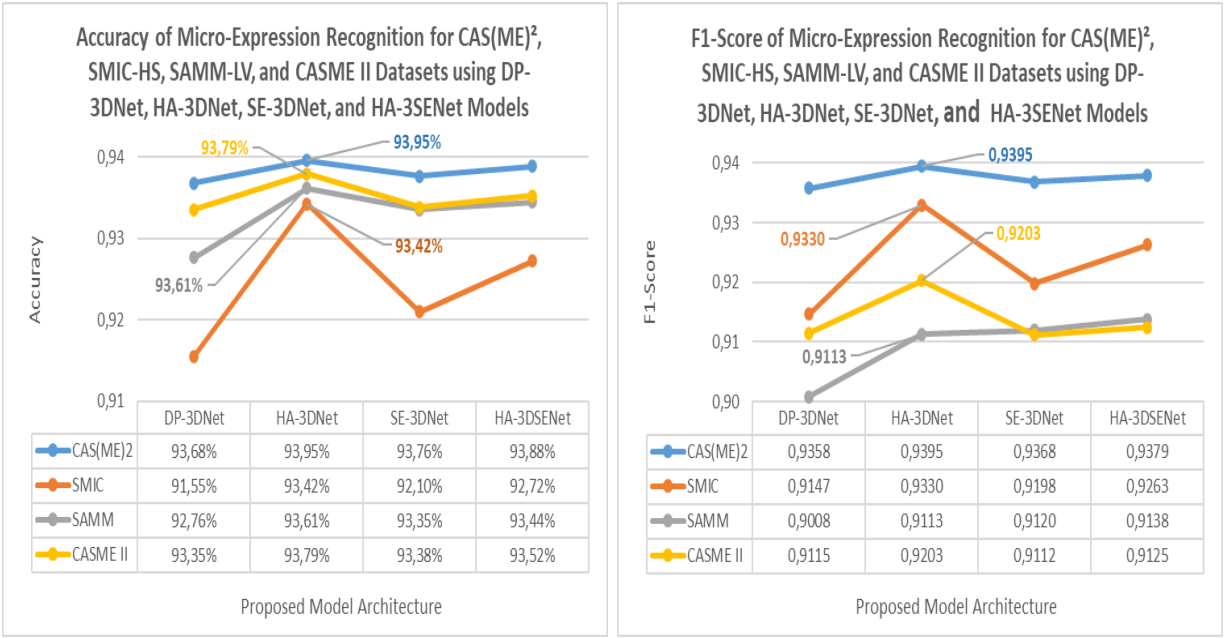

Figure 10. Accuracy and F1-score graphs for micro-expression recognition using CAS(ME)², SMIC, SAMM, CASME II datasets and DP-3DNet, HA-3DNet, SE-3DNet, HA-3SENet models.

When comparing models, Hybrid Attention-3DNet consistently outperforms other models across all datasets, indicating that incorporating spatial and temporal attention mechanisms substantially enhances the model's capacity for recognizing micro-expressions. This improvement highlights the importance of capturing spatial and temporal dependencies, crucial for identifying subtle and rapid facial expressions.

Hybrid Attention, with its combined spatial and temporal attention mechanisms, effectively enhances

micro-expression recognition accuracy. Our experiments demonstrate that Hybrid Attention improves accuracy and F1-score, particularly on the CAS(ME)² and SMIC datasets. This underscores the model's ability to capture expressions' complex spatial and temporal characteristics.

The spatial and temporal attention combination enables the model to detect subtle changes in expressions within very short durations, often challenging to identify in micro-expression videos. By focusing on critical facial features and temporal variations, Hybrid Attention improves the model's ability to differentiate emotions that appear briefly, yet convey meaningful information. Thus, incorporating Hybrid Attention significantly enhances micro-expression recognition, especially for fleeting expressions commonly found in micro-expressions.

Regarding the attention mechanism, the Hybrid Attention-3DNet architecture proves superior to Squeeze-and-Excitation-3DNet, suggesting that attention mechanisms offer more significant benefits in this context. This effectiveness likely stems from the specific nature of micro-expression data, where capturing temporal dynamics is essential. Although a combined model architecture with Hybrid Attention and Squeeze-and-Excitation was tested, results showed no substantial improvement over Hybrid Attention-3DNet alone, indicating that the main performance gain originates from the attention mechanism.

Dataset-performance variability across datasets indicates that, although Hybrid Attention-3DNet is highly reliable, dataset characteristics still influence its performance. The model consistently has high accuracy and F1 scores across datasets, but reflects good generalization capabilities. Additionally, pre-processing steps, such as data augmentation and class balancing, play an essential role**.** Data augmentation improves generalization by diversifying training samples, while class balancing prevents bias toward majority classes.

A visualization analysis was conducted on correct and incorrect classification results for specific micro-expressions, including "anger" and "fear," which often experience misclassification due to similar spatial and temporal patterns. This visualization highlights particular facial areas, such as the region around the eyes, that frequently cause misclassifications due to similar muscle movements in both expressions.

A case study on the "happiness" expression, which the model recognizes relatively quickly, was also performed. This recognition can be attributed to consistent spatial patterns around the mouth, aiding the model in differentiating this expression from others. This qualitative analysis provides insights into the model's strengths and weaknesses, particularly concerning subtle variations in micro-expressions.

For dataset performance, Hybrid Attention-3DNet achieved the highest accuracy with the SMIC dataset, likely due to multiple factors. SMIC has the largest sample size (about 164 video clips), enabling the model to learn and generalize patterns more effectively. Moreover, sample durations in SMIC vary significantly (from 9 to 343 seconds), allowing the model to capture spatial and temporal features. With a frame rate of 100 fps, the second highest among the datasets, SMIC provides ample spatial and temporal information, which is critical for classification. Its 640x480 resolution balances spatial detail with manageable data size.

Error analysis helps understand where and why the model makes mistakes. Based on the confusion matrix, some common errors in the CAS(ME)² dataset can be observed: for the Angry class, 7 Angry samples were classified as Disgust and 3 Angry samples were classified as Happy. For the Disgust class, 9 Disgust samples were classified as Angry and 5 Disgust samples were classified as Happy. For the Happy class, 6 Happy samples were classified as Angry and 7 Happy samples were classified as Disgust. This indicates confusion between particular classes, particularly between Angry and Disgust and between Happy and Disgust. A similar analysis for the other datasets can be found in Table 3.

Table 3. Accuracy, F1-Score and Error Rate of the Hybrid Attention – 3DNet model with CAS(ME)², SMIC, SAMM and CASME II Datasets.

Dataset	Accuracy (%)	F1-Score	Error Rate (%)
CAS(ME)^2	93.95	0.9395	6.05
SMIC	93.42	0.9330	5.58
SAMM	93.61	0.9113	6.39
CASME II	93.79	0,9203	6.21

In real-world applications, the developed micro-expression recognition model holds potential for various fields, including security, clinical psychology and human-computer interaction. For instance, accurately recognizing expressions of "fear" or "anger" can be crucial for detecting potential threats or high-stress situations in security settings. However, the model's limitations in differentiating between similar micro-expressions, such as "fear" and "anger," could pose challenges in these applications, as misclassifying these expressions may impact critical decisions.

In clinical psychology, recognizing more apparent expressions like "happiness" or "sadness" offers applications for non-invasively assessing patients' emotional states. The model can assist in evaluating emotional responses to specific stimuli; however, subtle variations in micro-expressions could be missed, particularly when spatial-temporal patterns overlap between expressions of interest.

Another limitation in practical deployment is the model's sensitivity to the characteristics of the training dataset. Variations in lighting, facial angles or environmental conditions may affect the model's performance outside controlled laboratory settings. Additional data augmentation and adjustments for varying lighting conditions could be considered in the implementation phase, enhancing the model's reliability across diverse real-world situations. This discussion underscores the importance of further optimizing the model and conducting additional testing under real-world conditions to enhance its reliability in practical applications. It also highlights future development opportunities focused on adapting the model to a broader range of scenarios.

7. COMPARISON WITH PREVIOUS WORKS

Tables 4, 5, 6 and 7 compare the accuracy and F1-Score between the latest and proposed approaches using datasets including CAS(ME)², SMIC, SAMM and CASME II. These tables show that the proposed approach performs relatively better than the state-of-the-art methods. Enhancing the spatial-temporal feature weight attention in the 3D dual-path convolutional network model using hybrid attention and the Squeeze-and-Excitation Block improved the evaluation metrics for accuracy and F1-score.

Table 4. Comparison of accuracy and F1-score between the proposed method and state-of-the-art models on the CAS(ME)² dataset.

Year	Methods	Accuracy (%)	F1-score
2021	MSFME-IR [21]	-	0.8103
2021	RMER-3DCNN [9]	79.31	-
2021	LEARNET [22]	76.33	-
2022	MERASTC [23]	91.20	0.9070
2022	Deep3DCANN [24]	90.00	0.8800
2022	SE-DenseNet-T+EVM [25]	92.96	0.9289
2023	MER-DBNN [10]	-	0.8103
2024	Dual Path-3DNet	93.68	0.9358
2024	**Hybrid Attention-3DNet**	**93.95**	**0.9395**
2024	Squeeze-and-Excitation-3DNet	93.76	0.9368
2024	Hybrid Attention-3DSENet	93.88	0.9379

Note: '-' indicates that the data was not available in the referenced study.

Table 5. Comparison of accuracy and F1-score between the proposed method and state-of-the-art models on the SMIC dataset.

Year	Methods	Accuracy (%)	F1-score
2021	RMER-3DCNN [9]	76.92	-
2022	3DCNN-MED [26]	80.94	-
2022	MERASTC [23]	79.30	0.7900
2023	MER-DBNN [10]	-	0.6687
2023	BDCN [27]	-	0.7859
2023	RNAS MER [28]	-	0.7443
2023	FRL-DGT [29]	-	0.749
2023	DS-3DCNN [30]	78.78	0.7887
2024	Dual Path-3DNet	91.55	0.9147
2024	**Hybrid Attention-3DNet**	**93.42**	**0.9330**
2024	Squeeze-and-Excitation-3DNet	92.10	0.9198
2024	Hybrid Attention-3DSENet	92.72	0.9263

Table 6. Comparison of accuracy and F1-score between the proposed method and state-of-the-art models on the SAMM dataset.

Year	Methods	Accuracy (%)	F1-score
2021	RMER-3DCNN [9]	73.91	-
2022	MERASTC [23]	83.80	0.8440
2022	Deep3DCANN [24]	93.00	0.8900
2023	DBMNet [11]	-	0.6494
2023	BDCN [27]	-	0.8538
2023	RNAS MER [28]	-	0.7880
2023	ADMME [31]	81.43	0.8161
2023	FRL-DGT [29]	-	0.7580
2023	DS-3DCNN [30]	79.17	0.7156
2024	Dual Path-3DNet	92.76	0.9008
2024	**Hybrid Attention-3DNet**	**93.61**	**0.9113**
2024	Squeeze-and-Excitation-3DNet	93.35	0.9120
2024	Hybrid Attention-3DSENet	93.44	0.9138

Table 7. Comparison of accuracy and F1-Score between the proposed method and state-of-the-art models on the CASME II dataset

Year	Methods	Accuracy (%)	F1-score
2022	MERASTC [23]	85.40	0.8620
2022	Deep3DCANN [24]	86.00	0.8400
2022	SE-DenseNet-T+EVM [25]	82.74	0.7659
2023	DBMNet [11]	-	0.6653
2023	MER-DBNN [10]	-	0.8189
2023	BDCN [27]	-	0.9501
2023	STCPNet [12]	91.46	0.8977
2023	RNAS MER [28]	-	0.8985
2023	ADMME [31]	86.34	0.8635
2023	FRL-DGT [29]	-	0.9030
2024	Dual Path-3DNet	93.35	0.9115
2024	**Hybrid Attention-3DNet**	**93.79**	**0.9203**
2024	Squeeze-and-Excitation-3DNet	93.38	0.9112
2024	Hybrid Attention-3DSENet	93.52	0.9125

8. CONCLUTION

This study developed a pipeline with multiple processing stages to recognize spontaneous micro-expressions effectively. The results indicate that the proposed method surpasses state-of-the-art approaches, achieving accuracy and F1-score values of 93.95% and 0.9395 on the CAS(ME)² dataset, 93.42% and 0.9330 on SMIC, 93.61% and 0.9113 on SAMM and 93.79% and 0.9203 on CASME II. Among the datasets, the SMIC dataset exhibited the lowest error rate at 5.58%, followed by CAS(ME)² at 6.05%, CASME II at 6.21% and SAMM with the highest error rate of 6.39%. The differences in accuracy and F1-score values can be attributed to the distinct characteristics of each dataset, even when the same pipeline is applied. This study highlights that the implemented pipeline has successfully enhanced micro-expression recognition accuracy, primarily due to the improved attention to spatial-temporal feature weights.

ACKNOWLEDGEMENTS

The first author is a dedicated Telkom Foundation of Education employee, serving as a lecturer at the School of Electrical Engineering, Telkom University. He is advancing his academic career by pursuing a doctoral program at the School of Electrical Engineering and Informatics, Bandung Institute of Technology. Telkom University strongly supports this study, reflecting the institution's commitment to fostering academic growth and contributing to advancements in electrical engineering and informatics. The author's ongoing studies and this study project demonstrate a synergy between professional responsibilities and academic pursuits, aiming to contribute significantly to academia and industry.

REFERENCES

[1] P. Zhang, X. Ben, R. Yan, C. Wu and C. Guo, "Micro-expression Recognition System," Optik (Stuttgart), vol. 127, no. 3, pp. 1395–1400, DOI: 10.1016/j.ijleo.2015.10.217, 2016.

[2] G. Zhao and X. Li, "Automatic Micro-expression Analysis: Open Challenges," Frontiers in Psychology, vol. 10, no. AUG, pp. 1–4, DOI: 10.3389/fpsyg.2019.01833, 2019.

[3] L. Zhou, X. Shao and Q. Mao, "A Survey of Micro-expression Recognition," Image and Vision Computing, vol. 105, p. 104043, DOI: 10.1016/j.imavis.2020.104043, 2021.

[4] Y. He, S. J. Wang, J. Li and M. H. Yap, "Spotting Macro-and Micro-expression Intervals in Long Video Sequences," Proc. of the 2020 15th IEEE Int. Conf. Autom. Face Gesture Recognition (FG 2020), pp. 742–748, DOI: 10.1109/FG47880.2020.00036, 2020.

[5] F. Qu, S. J. Wang, W. J. Yan, H. Li, S. Wu and X. Fu, "CAS(ME)^2): A Database for Spontaneous Macro-expression and Micro-expression Spotting and Recognition," IEEE Transactions on Affective Computing, vol. 9, no. 4, pp. 424–436, DOI: 10.1109/TAFFC.2017.2654440, 2018.

[6] X. Li, P. Tomas, H. Xiaohua, Z. Guoying and P. Matti, "A Spontaneous Micro-expression Database: Inducement, Collection and Baseline," Proc. of the 2013 10th IEEE Int. Conf. and Workshops on Automatic Face and Gesture Recognition (FG), DOI: 10.1109/FG.2013.6553717, Shanghai, China, 2013.

[7] A. K. Davison et al., "SAMM : A Spontaneous Micro-facial Movement Dataset," IEEE Transactions on Affective Computing, vol. 9, no. 1, pp. 116–129, DOI: 10.1109/TAFFC.2016.2573832, 2016.

[8] W. J. Yan et al., "CASME II: An Improved Spontaneous Micro-expression Database and the Baseline Evaluation," PLoS One, vol. 9, no. 1, pp. 1–8, DOI: 10.1371/journal.pone.0086041, 2014.

[9] Y. Jiao, M. Jing, Y. Hu and K. Sun, "Research on a Micro-expression Recognition Algorithm Based on 3D-CNN," Proc. of the 2021 3rd Int. Conf. Intell. Control. Meas. Signal Process. Intell. Oil Field (ICMSP 2021), no. Icmsp, pp. 221–225, DOI: 10.1109/ICMSP53480.2021.9513351, 2021.

[10] F. Guowen and L. Xi, "Micro-expression Recognition Based on Dual Branch Neural Network," Proc. of the 2023 Int. Conf. Artif. Intell. Comput. Inf. Technol. (AICIT 2023), no. 2020, pp. 2–5, DOI: 10.1109/AICIT59054.2023.10278020, 2023.

[11] Q. Zhou, S. Liu, Y. Wang and J. Wang, "Divided Block Multi-scale Convolutional Network for Micro-expression Recognition," Proc. of the 2022 1st Int. Conf. Cyber-Energy Syst. Intell. Energy (ICCSIE 2022), pp. 1–5, DOI: 10.1109/ICCSIE55183.2023.10175242, 2023.

[12] Z. Shang, J. Liu and X. Li, "Micro-expression Recognition Based on Spatio-temporal Capsule Network," IEEE Access, vol. 11, no. January, pp. 13704–13713, DOI: 10.1109/ACCESS.2023.3242871, 2023.

[13] H. Tian, W. Gong, W. Li and Y. Qian, "PASTFNet: A Paralleled Attention Spatio-temporal Fusion Network for Micro-expression Recognition," Medical and Biological Engineering and Computing, DOI: 10.1007/s11517-024-03041-y, 2024.

[14] A. J. Rakesh Kumar and B. Bhanu, "Relational Edge-node Graph Attention Network for Classification of Micro-expressions," Proc. of the IEEE Computer Society Conf. on Computer Vision and Pattern Recognition, vol. 2023-June, pp. 5819–5828, DOI: 10.1109/CVPRW59228.2023.00618, 2023.

[15] F. Li, P. Nie, M. You, Z. Chen and G. Wang, "Triple-ATFME: Triple-branch Attention Fusion Network for Micro-expression Recognition," Arabian J. for Science and Eng., DOI: 10.1007/s13369-024-08973-z, 2024.

[16] H. Insan, S. S. Prasetiyowati and Y. Sibaroni, "SMOTE-LOF and Borderline-SMOTE Performance to Overcome Imbalanced Data and Outliers on Classification," Proc. of the 2023 3rd Int. Conf. Intell. Cybern. Technol. Appl., pp. 136–141, DOI: 10.1109/icicyta60173.2023.10428902, 2024.

[17] A. Sagoolmuang, "Power-weighted kNN Classification for Handling Class Imbalanced Problem," Proc. of the 2021 2nd Int. Conf. Big Data Anal. Pract. (IBDAP 2021), pp. 42–47, DOI: 10.1109/IBDAP52511.2021.9552164, 2021.

[18] B. Irawan, N. P. Utama, R. Munir and A. Purwarianti, "Improving the Accuracy of Facial Micro-expression Rrecognition: Spatio-temporal Deep Learning with Enhanced Data Augmentation and Class Balancing," Interdisciplinary Journal of Information, Knowledge, and Management, vol. 19, pp. 1–15, DOI: 10.28945/5386, 2024.

[19] Y. Zhou, H. Chen, J. Li, Y. Wu, J. Wu and L. Chen, "ST-Attn: Spatial-," Proc. of the IEEE Int. Conf. Data Min. Work. (ICDMW), vol. 2019-Novem, no. November, pp. 609–614, DOI: 10.1109/ICDMW.2019.00092, 2019.

[20] Y. S. Gan, S. E. Lien, Y. C. Chiang and S. T. Liong, "LAENet for Micro-expression Recognition," Visual Computer, vol. 40, no. 2, pp. 585–599, DOI: 10.1007/s00371-023-02803-3, 2024.

[21] P. Sharma, S. Coleman, P. Yogarajah, L. Taggart and P. Samarasinghe, "Magnifying Spontaneous Facial Micro Expressions for Improved Recognition," Proc. of the Int. Conf. Pattern Recognit., pp. 7930–7936, DOI: 10.1109/ICPR48806.2021.9412585, 2020.

[22] M. Verma, S. K. Vipparthi, G. Singh and S. Murala, "LEARNet: Dynamic Imaging Network for Micro Expression Recognition," IEEE Transactions on Image Processing, vol. 29, no. c, pp. 1618–1627, DOI: 10.1109/TIP.2019.2912358, 2020.

[23] P. Gupta, "MERASTC: Micro-expression Recognition Using Effective Feature Encodings and 2D Convolutional Neural Network," IEEE Transactions on Affective Computing, vol. 14, no. 2, pp. 1431–1441, DOI: 10.1109/TAFFC.2021.3061967, 2023.

[24] S. Thuseethan, S. Rajasegarar and J. Yearwood, "Deep3DCANN : A Deep 3DCNN-ANN Framework for Spontaneous Micro-expression Recognition," Information Sciences (Ny), vol. 630, no. November 2022, pp. 341–355, DOI: 10.1016/j.ins.2022.11.113, 2023.

[25] L. Cai, H. Li, W. Dong and H. Fang, "Micro-expression Recognition Using 3D DenseNet Fused Squeeze-and-excitation Networks," Applied Soft Computing, vol. 119, p. 108594, DOI: 10.1016/j.asoc.2022.108594, 2022.

[26] W. S. P. Bayu and A. Setyanto, "3D CNN for Micro Expression Detection," Proc. of the 5th Int. Conf. Inf. Commun. Technol. A New W. to Make AI Useful Everyone New Norm. Era (ICOIACT 2022), pp. 397–401, DOI: 10.1109/ICOIACT55506.2022.9972194, 2022.

[27] B. Chen, K. H. Liu, Y. Xu, Q. Q. Wu and J. F. Yao, "Block Division Convolutional Network with Implicit Deep Features Augmentation for Micro-expression Recognition," IEEE Transactions on Multimedia, vol. 25, pp. 1345–1358, DOI: 10.1109/TMM.2022.3141616, 2023.

[28] M. Verma, P. Lubal, S. K. Vipparthi and M. Abdel-Mottaleb, "RNAS-MER: A Refined Neural Architecture Search with Hybrid Spatiotemporal Operations for Micro-expression Recognition," Proc. of the 2023 IEEE Winter Conf. Appl. Comput. Vision (WACV 2023), pp. 4759–4768, DOI: 10.1109/WACV56688.2023.00475, 2023.

[29] Z. Zhai and J. Zhao, "Feature Representation Learning with Adaptive Displacement Generation and Transformer Fusion for Micro-expression Recognition," Proc. of the 2023 IEEE/CVF Conf. Comput. Vis. Pattern Recognit., pp. 22086–22095, DOI: 10.1109/CVPR52729.2023.02115, 2023.

[30] Z. Li, Y. Zhang, H. Xing and K.-L. Chan, "Facial Micro-expression Recognition Using Double-Stream 3D Convolutional Neural Network with Domain Adaptation," Sensors, vol. 23, no. 7, DOI: 10.3390/s23073577, 2023.

[31] Y. Wang, H. U. Shi and R. Wang, "Action Decouple Multi-tasking for Micro-expression Recognition," IEEE Access, vol. 11, no. June, pp. 82978–82988, DOI: 10.1109/ACCESS.2023.3301950, 2023.

ملخص البحث:

تقترح هذه الورقة البحثية نموذجاً مبتكراً للتعرّف إلى تعبيرات الوجْه الدّقيقة باستخدام شبكةٍ عصبيةٍ التفافية محسّنة عن طريق الاهتمام الهجين وكُتَل الانضغاط والإثارة. وتتلخّص الأهداف الرئيسية للنّموذج في: (1) تحسين استخلاص السِّمات باستخدام بِنية الشّبكة العصبية الالتفافية، (2) تحسين تمثيل البيانات عبر زيادة الصُّوَر على نحوٍ هادفٍ وتوزيع سِمات الصُّوَر بشكلٍ متوازنٍ، (3) تحسين انْدماج السِّمات باستخدام تقنيات الشّبكات الدّارجة في أدبيات الموضوع.

لقد تمّ إجراء تجارب للنّموذج المقترح على عددٍ من مجموعات البيانات. وقد برهن نموذج الشّبكة ثلاثية الأبعاد مع الاهتمام الهجين على قيمٍ متفوّقةٍ للدّقّة على جميع مجموعات البيانات المستخدمة مقارنةً بالنّماذج المماثلة المستخدمة في دراساتٍ سابقة أخرى على مجموعات البيانات ذاتها. وتدلُّ النّتائج الّتي تمّ الحصول عليها على أنّ النّموذج المقترح يمتاز بالفاعلية والمتانة، مع إمكانية استخدامه في مدى واسعٍ من الاستخدامات المتعلّقة بتمييز التّعبيرات الدّقيقة للوجْه، مثل الغَضَب والخَوف والسّعادة وغيرها.