© JUL 2026 | IRE Journals | Volume 10 Issue 1 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV10I1-1719374

Development of a Lightweight Temporal Convolutional Network for Micro Expression Recognition

ADEYEMI, IFEOLUWA OLAMIJI[1], FALOHUN, ADELEYE SAMUEL[2], OGUNTOYE JONATHAN PONMILE[3], AYINLA, MICHAEL OLUWASEUN[4]
[1,2 3,4]*Department of Computer Engineering, Ladoke Akintola University of Technology, Ogbomoso, Oyo State, Nigeria*

Abstract- ***Microexpression recognition (MER) remains challenging because facial movements are short, subtle, and usually require computationally expensive temporal models. This paper presents a lightweight depthwise-separable Temporal Convolutional Network (DS-TCN) for MER on the CASME II dataset. The model uses a dual-stream design that combines grayscale facial-frame features with optical-flow motion features after face cropping, Eulerian video magnification, resizing, and SSIM-based key-frame selection. Standard temporal convolutions are replaced with depthwise separable temporal convolutions to reduce parameter count and floating-point operations while preserving sequence learning. Three-fold cross-validation produced a validation accuracy of 72.6% and a macro-F1 score of 67.58%. Compared with the standard TCN baseline, the proposed DS-TCN reduced parameters from 1.57 M to 0.219 M, FLOPs from 10.33 G to 1.50 G, and CPU inference time from 580 ms to 210 ms. The results show that lightweight temporal modeling can provide competitive MER performance with substantially lower computational cost.***

Index Terms- ***CASME II, Depthwise Separable Convolution, Microexpression Recognition, Optical Flow, Temporal Convolutional Network***

I. INTRODUCTION

Microexpressions are brief and involuntary facial movements that often last less than half a second and may reveal concealed emotions. Their short duration and low movement intensity make automatic recognition difficult, especially when models depend on static appearance features or require large training sets. These challenges are important because MER has practical value in human-computer interaction, healthcare monitoring, security screening, and affective computing.

Earlier MER methods relied largely on handcrafted descriptors and selected key frames. Although such techniques captured visible texture changes, they did not fully model the temporal transitions that define microexpressions. Recent deep learning methods have improved recognition through optical flow, 3D convolution, recurrent networks, graph convolution, and attention mechanisms. However, improved recognition often comes with higher model depth, larger parameter counts, and slower inference, limiting deployment on CPU-only and resource-constrained devices.

This study addresses the efficiency problem by developing a lightweight Temporal Convolutional Network in which standard temporal convolutions are replaced with depthwise separable convolutions. The contribution of the study is threefold: first, it proposes a dual-stream DS-TCN that integrates spatial facial cues and optical-flow temporal features; second, it provides empirical evidence from CASME II using three-fold cross-validation; and third, it evaluates both recognition performance and computational efficiency using accuracy, macro-F1 score, parameters, FLOPs, and CPU inference time.

A. Contribution of the study

The main concern addressed in this paper is therefore the trade-off between recognition performance and computational economy. A model that performs well but requires a high number of parameters, large FLOPs, or slow CPU inference is less useful for field deployment, especially where microexpression recognition is expected to run on laptops, mobile devices, embedded cameras, or edge-computing platforms. The proposed approach is positioned within this constraint: it does not attempt to increase depth or add heavy attention blocks merely to

improve accuracy. Instead, it retains temporal sequence learning while compressing the convolutional operations responsible for most of the computational load.

The contribution of the paper is threefold. First, it presents a dual-stream pipeline that combines grayscale facial appearance information with optical-flow motion information so that both static and dynamic facial cues are retained. Second, it redesigns the temporal convolutional stage by replacing standard temporal convolution with depthwise separable temporal convolution, supported by dilation and residual learning. Third, it evaluates the model using both recognition metrics and deployment-oriented efficiency metrics, including trainable parameters, FLOPs, and CPU inference time. This combination makes the study useful not only as a recognition experiment but also as an efficiency-oriented MER design for practical implementation.

II. RELATED WORK

MER research has progressed from handcrafted and spatial descriptors toward spatiotemporal deep learning. Motion-based techniques such as optical flow are useful because they encode the direction and magnitude of subtle facial movement. CNN-based and recurrent designs further improve feature learning, while graph-based and temporal convolutional approaches capture inter-region and sequence dependencies. Despite these gains, many recent models remain expensive because they use 3D convolution, multi-scale blocks, attention modules, or graph operations.

Lightweight neural design has therefore become an important direction in MER. Compact networks reduce memory use and inference latency, but many lightweight MER studies focus more on spatial compression than on efficient temporal sequence modeling. Depthwise separable convolution provides a suitable route for reducing temporal convolution cost because it separates channel-wise filtering from pointwise feature combination. The present work applies this idea directly to TCN-based MER, preserving temporal receptive fields through dilation and residual connections while reducing computational burden.

A. Handcrafted, temporal and lightweight MER directions

Earlier MER systems commonly relied on handcrafted descriptors such as local binary patterns, optical flow histograms, and local motion descriptors. These methods were useful because they could operate on small datasets, but their performance depended heavily on manually designed features and carefully selected facial regions. Later CNN-based models improved spatial feature learning by allowing the network to extract discriminative facial patterns directly from images or optical-flow maps. However, many CNN-based methods still treated the expression as a frame-level or apex-frame problem and therefore did not always model how the expression evolves from onset to apex and offset.

Temporal models address this weakness by learning motion continuity across the video sequence. Recurrent networks, 3D CNNs, graph convolutional networks, and temporal convolutional networks have all been used to capture the short-lived dynamics of microexpressions. Their advantage is that they can detect subtle changes that are not visible in a single frame. Their disadvantage is that the extra temporal processing often increases model size, memory usage, and computation time. This problem is particularly important for MER because microexpression datasets are small, highly imbalanced, and difficult to annotate, so a very large model can overfit or become unsuitable for real-time deployment.

Lightweight MER has consequently become an important research direction. Compact architectures, shallow attention designs, mobile networks, and separable convolution blocks have been proposed to reduce cost while preserving useful discriminative power. The remaining gap, however, is that many lightweight MER designs are more spatial than temporal, while several temporal models remain computationally expensive. This study focuses on that gap by retaining the TCN structure for sequence modelling but compressing its convolutional operations using depthwise separable convolution.

© JUL 2026 | IRE Journals | Volume 10 Issue 1 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV10I1-1719374

Table 1 summarizes representative studies and the limitations that motivated the present work.

Table 1
Summary of relevant MER methods and remaining gaps

Study	Method	Dataset	Reported result	Main limitation
Li et al. (2018)	3D flow-based CNN	SMIC, CASME, CASME II	72.4%, 74.1%, 73.6% accuracy	Limited cross-dataset generalization
Zhu & Chen (2020)	Dual-stream temporal CNN	CASME II, SMIC	75.6%, 74.2% accuracy	Limited dataset diversity
Verma et al. (2021)	AutoMER	CASME II, SMIC, SAMM	81.3%, 79.6%, 80.5% accuracy	Higher computational demand
Cai et al. (2023)	GCN + LSTM	MMEW, SAMM	77.8%, 76.3% accuracy	Heavy spatiotemporal design
Jin et al. (2024)	Multi-scale 3D ResNet	SMIC, SAMM, CASME II	74.6%, 84.77%, 91.35% accuracy	High model complexity
Yu et al. (2024)	Shallow attention network	SMIC, CASME II, SAMM	0.278 M parameters	Less focused on TCN-based sequence learning
Xue et al. (2025)	Region-focused lightweight network	SAMM, CASME II, SMIC	85.31% accuracy; 75.3 FPS	Uses region-attention design rather than lightweight TCN

Note. The comparison highlights the need for an MER model that combines explicit temporal learning with low computational cost.

III. METHODOLOGY

The proposed system used a dual-stream DS-TCN architecture. The spatial stream learned facial appearance features from grayscale frames, while the temporal stream learned motion information from dense optical flow. The two feature embeddings were projected into a common feature space, fused with a small attention-based module, and passed through depthwise-separable temporal convolutional blocks with dilation factors of 1, 2, 4, and 8.

The CASME II dataset was used because it provides high-frame-rate spontaneous microexpression clips. The clips were merged into four classes to reduce class sparsity: negative, positive, surprise, and others. This grouping was necessary because some original labels, such as fear and sadness, contained very few samples. The final class distribution is reported in Table 2.

A. Dataset preparation and label merging

The CASME II database was selected because it records spontaneous microexpressions at a high frame rate and provides frame-level annotations that are suitable for temporal modelling. Since some original emotion categories contain very few samples, the labels were merged into four categories: negative, positive, surprise, and others. This merging strategy reduces extreme class sparsity and creates a more stable classification setting for three-fold cross-validation. The merged distribution still remains imbalanced, with the others class having the largest number of samples and the surprise class having the fewest samples. For this reason, accuracy alone was not treated as sufficient; macro-F1 was also reported to reflect class-level balance.

The preprocessing pipeline was designed to reduce irrelevant variation before temporal learning. Face detection and cropping removed background, hair, clothing, and other non-facial content. Eulerian video magnification was then used to amplify subtle muscular changes that could otherwise be missed by a lightweight network. The extracted facial regions were resized to 128 × 128 grayscale frames to balance visible facial detail with computational economy. Finally, SSIM-based key-frame selection was used to prioritize frames with greater structural change, thereby reducing redundant information and focusing the model on informative temporal segments.

This preprocessing design is important because microexpression signals are easily obscured by irrelevant appearance variation. If background pixels, head movement, or redundant frames are passed directly into the temporal model, the network may learn non-emotional patterns rather than the short muscular changes that define the expression. Cropping and resizing therefore standardize the input, while motion magnification makes the subtle deformation more visible to the feature extractor. The SSIM selection stage further reduces temporal redundancy by giving priority to frames that contain meaningful structural difference from neighbouring frames.

The four-class grouping also supports a fairer evaluation of the lightweight model. The original CASME II categories are not equally represented, and training a compact temporal model on very sparse labels may produce unstable class boundaries. By merging related emotion labels, the experiment retains clinically and computationally meaningful categories while reducing the risk that the classifier simply memorizes extremely rare labels. This decision is reflected in the use of macro-F1, because the model must still perform across all final categories rather than only the dominant class.

Each clip was treated as a short temporal sequence rather than as an isolated apex-frame image. This decision is consistent with the purpose of using a TCN because the model must learn how the facial state changes across time. In practice, the sequence representation allows the spatial stream to retain frame-level appearance while the temporal stream supplies motion cues derived from optical flow. The two streams therefore compensate for one another: appearance features provide facial context, while motion features emphasize the subtle direction and magnitude of muscular displacement. This dual representation is especially useful when a visible texture change is weak but the optical-flow field still shows coordinated facial motion.

The implementation also considered computational cost at every preprocessing and modelling stage. Grayscale conversion reduces input channels, the 128 × 128 resolution limits unnecessary pixel operations, and key-frame selection lowers redundancy before temporal modelling. These choices are not merely convenience steps; they are part of the lightweight design philosophy of the paper. A compact model can only remain efficient if the input pipeline is also controlled. The dataset preparation stage therefore supports the model architecture by reducing avoidable computation before the frames reach the DS-TCN blocks.

The preprocessing output was organized into fixed-length 64-frame sequences. This fixed sequence length makes batching consistent and allows the temporal convolution blocks to receive a stable input shape during training. Where a clip contained more frames than required, the selected sequence emphasized the most informative motion interval; where the available frames were fewer, the sequence was standardized so that the model could still process it without architectural changes. This practical normalization step is essential for small MER datasets because clips vary in duration and expression phases.

Table 2
CASME II emotion categories after merging

Final category	Original components	Counts per emotion	Final count
Others	Others, repression	99 + 27	126
Negative	Disgust, sadness, fear	63 + 7 + 2	72

© JUL 2026 | IRE Journals | Volume 10 Issue 1 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV10I1-1719374

Positive	Happiness	32	32
Surprise	Surprise	25	25
Total	-	-	255

Note. Categories were merged to reduce severe class imbalance in the original CASME II labels.

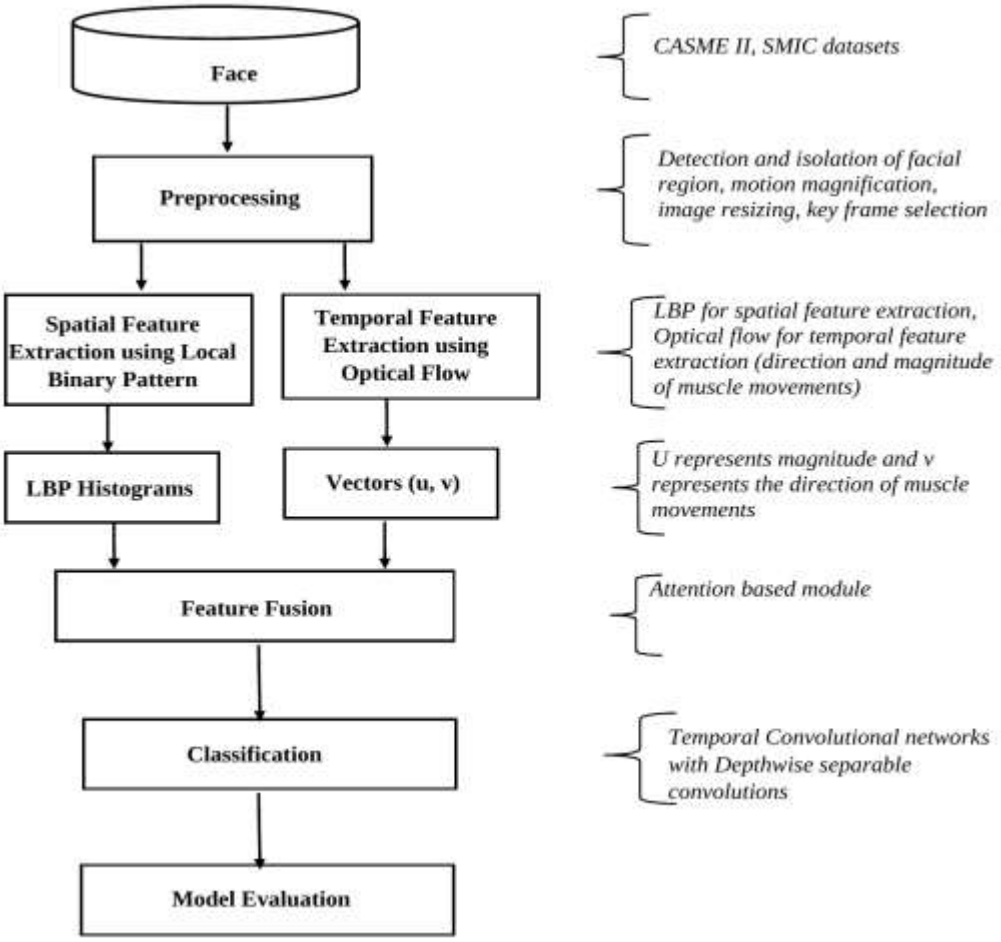

Figure 1: Scheme for the microexpression recognition system

Preprocessing and feature extraction

Preprocessing involved face detection and cropping, Eulerian video magnification, resizing, and key-frame selection. The Viola-Jones detector isolated the face region to reduce background noise. Eulerian Video Magnification amplified subtle movement cues before frames were resized to 128 × 128 pixels. SSIM was used to select informative key frames because lower structural similarity indicates larger facial change across frames.

Dense Farnebäck optical flow was computed between consecutive frames to represent horizontal and vertical motion components. The temporal stream therefore captured short movement trajectories, while the spatial stream supplied facial appearance cues. Batch normalization, ReLU activation, global average pooling, dropout, and residual connections were used to stabilize learning and reduce overfitting.

B. Feature extraction and DS-TCN formulation

The feature extraction stage used two complementary streams. The spatial stream processed grayscale facial frames and learned appearance cues such as local facial texture, cheek movement, eyebrow displacement, lip tightening, and eye-region changes. The temporal stream processed dense Farnebäck optical flow between consecutive frames, allowing the model to represent horizontal and vertical facial

motion. Optical flow was important because microexpressions are often too subtle to be recognized from appearance alone; the direction and magnitude of movement provide an additional signal that improves recognition of brief emotional leakage.
The DS-TCN architecture projected the spatial and temporal features into a common 64-dimensional embedding space and fused them through a small attention-based module. The fused sequence was then passed through temporal convolution blocks with dilation factors of 1, 2, 4, and 8. Dilation expanded the temporal receptive field without adding many layers, while residual connections improved gradient flow and stabilized learning. In the standard TCN, a temporal convolution with kernel size k, input channels Cin, and output channels Cout has parameter cost kCinCout. In the proposed DS-TCN, this operation is decomposed into a depthwise stage and a pointwise stage, giving a lower cost of kCin + CinCout. This reduction is the basis of the model compression reported in the results.

Training was designed for the small and imbalanced nature of the dataset. AdamW optimization was used with a One-Cycle learning-rate schedule, dropout was introduced in the temporal blocks, and early stopping was applied when validation loss failed to improve. Focal loss was selected because it reduces the dominance of easier majority-class examples and encourages the model to pay more attention to difficult or under-represented classes. The final protocol therefore combines architectural compression with training choices that reduce overfitting and class imbalance.

Table 3
Hyperparameter configuration

Hyperparameter	Configuration
Input image size	128 × 128 grayscale
Sequence length	64 frames
Batch size	4
Optimizer	AdamW
Initial learning rate	1e-3 with OneCycleLR peak
Loss function	Focal loss ($\gamma = 1.5$)
Training epochs	20 epochs with early stopping
Cross-validation	3-fold cross-validation
Attention fusion	MLP attention
Device	CPU
Model type	Two-stream DS-TCN

Note. The configuration follows the thesis experimental setup and was selected to support small-sample training and CPU-based inference.

IV. RESULTS AND DISCUSSION

The DS-TCN was evaluated using three-fold cross-validation. Across the folds, validation accuracy increased steadily during training and validation loss decreased progressively, showing stable learning. The final mean validation accuracy was 72.6%, while the macro-F1 score was 67.58%. The gap between accuracy and F1-score indicates that the model recognized larger classes more reliably than minority classes.

A. Evaluation emphasis

The results were interpreted from two perspectives. The first perspective concerns recognition quality, which was measured through validation accuracy, precision, recall, and macro-F1. The second perspective concerns computational efficiency, which was measured through parameter count, FLOPs, and CPU inference time. Reporting both perspectives is important because a lightweight MER system should not be judged only by recognition accuracy. A compact model must also demonstrate that the reduction in computation is large enough to justify its use in resource-constrained environments.

Three-fold cross-validation was used to reduce dependence on a single train-validation split. In each fold, the model was trained on two-thirds of the data and validated on the remaining one-third. The final values were reported as averages across the folds, allowing the learning behaviour to be observed more reliably. The validation graphs in Figures 2 and 3 therefore summarize not only one training run but the overall trend across the three-fold experimental protocol.

© JUL 2026 | IRE Journals | Volume 10 Issue 1 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV10I1-1719374

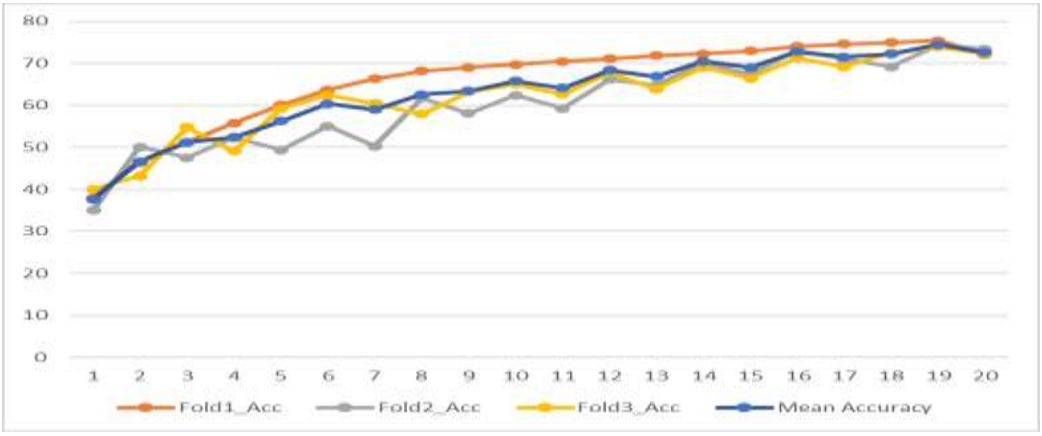

Figure 2: Validation accuracy across training epochs

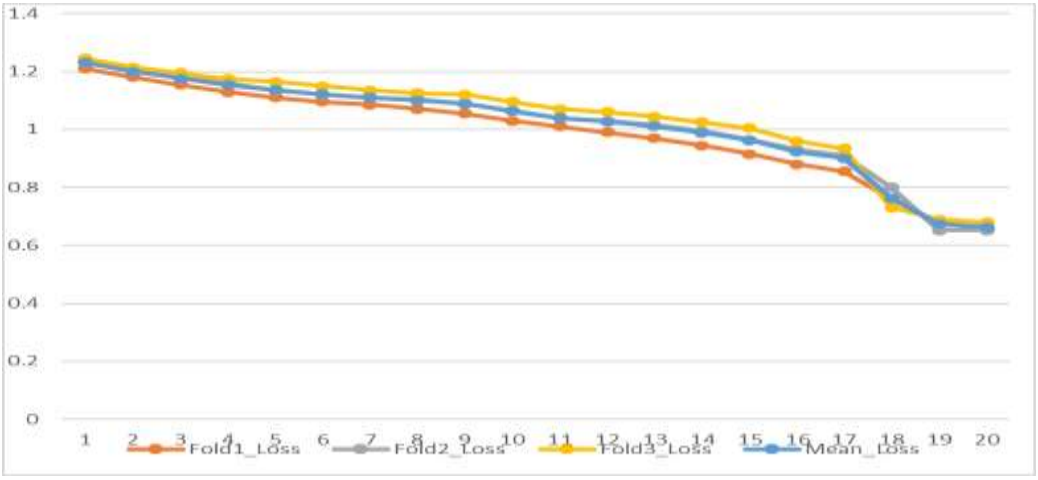

Figure 3: Validation loss across training epochs

A. Training behaviour and convergence

The training curves show that the proposed model learned progressively over the first part of the training cycle. Validation accuracy increased steadily before reaching a plateau, while validation loss declined and later stabilized. This pattern suggests that the DS-TCN extracted useful spatiotemporal features without requiring prolonged training. The stabilization of validation loss also supports the use of early stopping, since additional epochs after convergence produced limited gains and could increase the risk of overfitting on the small CASME II dataset.

The gap between validation accuracy and macro-F1 is also meaningful. The validation accuracy of 72.6% shows that the model classified a majority of samples correctly, while the macro-F1 score of 67.58% indicates that minority classes were harder to classify

consistently. This is expected in MER because the facial cues for some emotion categories overlap and the number of available training samples is uneven. The analysis therefore emphasizes both overall recognition and class-level reliability rather than presenting accuracy as the only measure of model performance.

© JUL 2026 | IRE Journals | Volume 10 Issue 1 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV10I1-1719374

Table 4
Fold-by-fold training and validation performance

Fold	Training accuracy (%)	Validation accuracy (%)	Training loss	Validation loss	Macro-F1 (%)
1	80.4	72.1	0.495	0.658	70.8
2	81.7	73.4	0.472	0.631	71.6
3	79.9	72.3	0.508	0.669	70.9
Mean ± SD	80.7 ± 0.9	72.6 ± 0.55	0.492 ± 0.018	0.653 ± 0.019	71.1 ± 0.35

Note. Mean and standard deviation are computed across the three reported folds.

Table 5
Class-wise precision, recall, and F1-score

Class	Precision (%)	Recall (%)	F1 (%)
Negative	69.23	75.00	72.00
Others	78.38	70.73	74.27
Positive	58.33	63.64	60.87
Surprise	60.00	66.67	63.16
Macro average	66.49	69.01	67.58

Note. The minority positive and surprise classes recorded lower F1-scores than the larger negative and others classes.

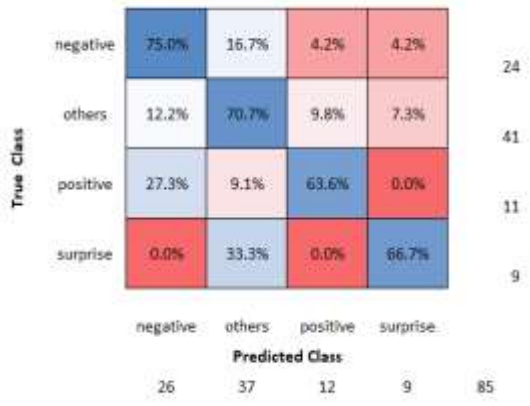

Figure 4: Confusion matrix for the DS-TCN model

The confusion matrix shows a strong diagonal pattern, indicating that most samples were correctly assigned to their classes. The major errors occurred between negative and others, which is expected because both groups may contain subtle lower-face movements such as lip tightening, brief frowning, and small tension patterns. Positive and surprise retained distinguishable cues, but their lower sample counts limited class-level balance.

Ablation and computational efficiency

The ablation analysis focuses on the main architectural change in the study: replacing standard temporal convolution with depthwise separable temporal convolution while keeping the dual-stream raw-frame and optical-flow design. This comparison is methodologically useful because it isolates the effect of the lightweight temporal block under the same general training setting.

B. Interpretation of class-wise performance

The class-wise results show that the model performed better on classes with more stable or more frequent training examples. The negative and others categories achieved stronger F1-scores than the positive and surprise categories. This result agrees with the merged class distribution, where positive and surprise had fewer samples. The confusion matrix further shows that most errors occurred between negative and others, which is reasonable because both categories may involve visually similar low-intensity movements such as brief brow lowering, lip pressing, or small lower-face tension. The positive and surprise categories showed more distinct cues in some instances, but their smaller sample sizes limited the stability of learning.

The ablation analysis isolates the central design decision of the paper: replacing standard temporal convolution with depthwise separable temporal convolution. The baseline TCN retained slightly higher recognition performance, but the gain was small compared with its computational cost. The DS-

TCN therefore represents a deliberate trade-off. It sacrifices a small amount of recognition capacity while achieving a large reduction in parameters, FLOPs, and CPU inference time. In practical MER systems, this trade-off is valuable because real-time operation and low-resource deployment may be more important than a marginal increase in validation accuracy.

Table 6
Ablation study: standard TCN versus depthwise-separable TCN

Model variant	Temporal convolution	Accuracy (%)	Macro-F1 (%)	Params (M)	FLOPs (G)	Inference time (ms)
Baseline TCN	Standard convolution	74.50	71.00	1.57	10.33	580
Proposed DS-TCN	Depthwise separable convolution	72.60	67.58	0.219	1.50	210
Change	DS-TCN minus baseline	-1.90 pp	-3.42 pp	-86.1%	-85.5%	-63.8%

Note. pp = percentage points. The comparison isolates the effect of replacing standard temporal convolution with depthwise separable temporal convolution.

Table 7
Component-level parameter ablation for DS-TCN and baseline TCN

Component	Description	DS-TCN	Baseline TCN
Spatial CNN extractor	Initial spatial feature extraction	2,000	2,000
TCN blocks (raw)	Temporal extractor for raw-frame stream	98,000	773,220
Projection head (raw)	1 × 1 channel alignment	4,100	4,100
Temporal CNN extractor	Processing optical-flow frames	2,000	2,000
TCN blocks (flow)	Temporal extractor for motion stream	98,000	773,220
Projection head (flow)	Dimensionality reduction	4,100	4,100
Fusion + attention	Feature fusion and MLP attention	9,000	9,000
Classifier	Fully connected classifier	2,200	2,200
Total	-	219,400 (0.219 M)	1,570,000 (1.57 M)

Note. The parameter reduction is concentrated in the temporal convolutional blocks.

Table 8
Component-level FLOPs ablation for DS-TCN and baseline TCN

Component	DS-TCN	Baseline TCN
CNN extractor (raw)	102 M	102 M
TCN blocks (raw)	640 M	5,050 M
Projection head (raw)	9 M	9 M
CNN extractor (flow)	102 M	102 M
TCN blocks (flow)	640 M	5,050 M
Projection head (flow)	9 M	9 M

Fusion + attention	3 M	3 M
Classifier	2 M	2 M
Total	1,507 M (1.50 G)	10,327 M (10.33 G)

Note. Most FLOP savings came from replacing standard TCN blocks with depthwise-separable temporal blocks.

Table 9
Model performance comparison with selected prior methods

Method	Author/year	Accuracy (%)	Macro-F1 (%)	Notes
LBP-TOP	Khor et al. (2011)	39.68	35.89	Handcrafted baseline
CNN+LSTM	Kim et al. (2016)	60.89	N/A	CNN + temporal modeling
Hierarchical STLBP-IP	Zong et al. (2019)	64.49	61.25	Statistical spatiotemporal
ME-Booster	Peng et al. (2019)	70.85	N/A	Motion amplification
DSSN	Khor et al. (2019)	71.16	72.97	Deep spatial network
SSSN	Khor et al. (2019)	71.19	71.51	Enhanced spatial network
Graph-TCN	Li et al. (2021)	73.98	72.46	Graph-structured TCN
Baseline TCN	Developed model	74.50	71.00	Standard temporal convolution
DS-TCN	Developed model	72.60	67.58	Lightweight temporal model

Note. The DS-TCN gives competitive accuracy while requiring substantially fewer parameters and FLOPs.

The efficiency gains are substantial. The proposed DS-TCN used approximately 14.5% of the FLOPs required by the standard TCN and reduced CPU inference time from 580 ms to 210 ms per sample. This confirms that the efficiency benefit is not only theoretical; it is reflected in actual runtime. The accuracy reduction of 1.90 percentage points relative to the standard TCN is modest when weighed against the large reduction in model size and computation.

Compared with existing MER methods, the DS-TCN performs better than older handcrafted and early recurrent models while remaining close to heavier temporal and graph-based networks. The main limitation is weaker class balance, especially for positive and surprise, which had fewer samples. This limitation is consistent with the small-sample and class-imbalance characteristics of public MER datasets.

C. Practical implication, limitations and future direction

The computational results indicate that the proposed model is more suitable for deployment than a standard TCN. With approximately 0.219 million parameters and 1.50 GFLOPs per sample, the DS-TCN places lower demand on memory and computation. Its CPU inference time of about 210 ms per sample also shows that the model can operate without relying on GPU-only execution. This makes it relevant for applications such as low-cost monitoring systems, human-computer interaction interfaces, and edge-based emotion analysis, where hardware resources are limited and latency matters.

Despite these strengths, the study has limitations. The experiment was conducted on CASME II, and although the dataset is widely used, it is still small and demographically limited. The use of three-fold cross-validation provides a reasonable estimate of performance, but stronger generalization claims would require testing on SAMM, SMIC, and composite MER datasets. The lower macro-F1 score also shows that minority classes remain difficult. Future work should therefore explore class-aware sampling, temporal augmentation, cross-dataset validation, and lightweight attention modules that improve minority-class performance without

reversing the computational savings achieved by depthwise separable temporal convolution.

Overall, the findings support the central argument of the paper: temporal modelling does not have to be computationally heavy to be useful for microexpression recognition. A carefully compressed TCN can retain most of the recognition benefit of temporal learning while reducing the model footprint enough to make deployment more realistic.

V. CONCLUSION

This paper presented a lightweight DS-TCN for microexpression recognition using CASME II. The model combines spatial grayscale-frame features and optical-flow temporal features, then applies depthwise-separable temporal convolution to reduce computational cost. The proposed model achieved 72.6% validation accuracy and 67.58% macro-F1 score while reducing parameters, FLOPs, and CPU inference time by large margins compared with a standard TCN. The findings show that temporal learning can be retained in a compact model suitable for CPU-only and resource-constrained MER applications.

Future work should evaluate the model on SAMM, SMIC, and composite MER datasets to test cross-dataset generalization. Further improvement may also come from class-aware augmentation, balanced sampling, lightweight adaptive attention, and subject-independent validation protocols such as leave-one-subject-out evaluation.

REFERENCES

[1] Aouayeb, M., Hamidouche, W., Soladié, C., Kpalma, K., & Séguier, R. (2021). Micro-expression recognition from local facial regions. Signal Processing: Image Communication, 99, 116457.

[2] Cai, X., Tang, H., & Chai, L. (2023). Micro expression recognition based on graph convolutional networks with LSTM. In 2023 35th Chinese Control and Decision Conference (CCDC) (pp. 5449-5453). IEEE.

[3] Gan, Y. S., Liong, S. T., Yau, W. C., Huang, Y. C., & Tan, L. K. (2019). OFF-ApexNet on micro-expression recognition system. Signal Processing: Image Communication, 74, 129-139.

[4] Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., & Adam, H. (2017). MobileNets: Efficient convolutional neural networks for mobile vision applications. arXiv:1704.04861.

[5] Jin, H., He, N., Li, Z., & Yang, P. (2024). Micro-expression recognition based on multi-scale 3D residual convolutional neural network. Mathematical Biosciences and Engineering, 21(4), 5007-5031.

[6] Khor, H. Q., See, J., Liong, S. T., Phan, R. C. W., & Lin, W. (2019). Dual-stream shallow networks for facial micro-expression recognition. In IEEE International Conference on Image Processing (pp. 36-40). IEEE.

[7] Kim, D. H., Baddar, W. J., & Ro, Y. M. (2016). Micro-expression recognition with expression-state constrained spatio-temporal feature representations. In ACM International Conference on Multimedia (pp. 382-386).

[8] Li, J., Wang, Y., See, J., & Liu, W. (2018). Micro-expression recognition based on 3D flow convolutional neural network. Pattern Analysis and Applications, 22, 1331-1339.

[9] Peng, W., Hong, X., Xu, Y., & Zhao, G. (2019). A boost in revealing subtle facial expressions: A consolidated Eulerian framework. In IEEE FG 2019 (pp. 1-5).

[10] Verma, M., Satish, M., Reddy, K., Meedimale, Y., & Kumar, S. (2021). AutoMER: Spatiotemporal neural architecture search for microexpression recognition. IEEE Transactions on Affective Computing.

[11] Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. (2004). Image quality assessment: From error visibility to structural similarity. IEEE Transactions on Image Processing, 13(4), 600-612.

[12] Wu, H.-Y., Rubinstein, M., Shih, E., Guttag, J., Durand, F., & Freeman, W. (2012). Eulerian video magnification for revealing subtle changes in the world. ACM Transactions on Graphics, 31(4), 1-8.

[13] Xue, F., Zhang, Y., Shao, Z., Cui, W., & Yuan, Z. (2025). HIM-PyraNet: Hierarchical attention and region-focused lightweight network for micro-expression recognition. Concurrency and Computation: Practice and Experience.

[14] Yu, Z., Chen, X., & Qu, C. (2024). SDGSA: A lightweight shallow dual-group symmetric attention network for micro-expression recognition. Complex & Intelligent Systems, 10, 8143-8162.

[15] Zhang, F., Huang, Z., Zhang, X., & Jin, Q. (2024). Adaptive temporal motion guided graph convolution network for micro-expression recognition. In IEEE International Conference on Multimedia and Expo.

[16] Zhang, L., & Arandjelović, O. (2021). Review of automatic microexpression recognition in the past decade. Machine Learning and Knowledge Extraction, 3(2), 414-434.

[17] Zong, Y., Zhang, M., & Zhao, G. (2020). Toward bridging microexpressions from different domains via domain adaptation methods. IEEE Transactions on Affective Computing, 13, 1117-1127.