Grenze International Journal of Engineering and Technology, June Issue

Micro-Expression Recognition based on TV-L1 based Hierarchical Transformer Network

Savitha N[1], Mohith Gowda K[2], T S Abhiram Hatawar[3], Goutam N T[4] and R V Niharika[5]
[1-5]JSS Science and Technology University Mysuru, India
Email: savithan@jssstuniv.in, mohithgowdak.kote@gmail.com, abhiramhatwar2003@gmail.com, goutamt2504@gmail.com, niharikarao2103@gmail.com

***Abstract*— Accurate human emotion and micro-expression interpretation is essential to enhance human-computer interaction, security system improvement, and supporting psychological studies. Micro-expression analysis enables the system to identify subtle non-verbal behavior, thus enabling empathetic and intuitive communication in security, customer service, and mental illness detection. This paper suggests a novel approach of micro-expression recognition by integrating a Hierarchical Transformer Network (HTNet) and the TV-L1 optical flow method. HTNet hierarchical transformer network efficiently extracts and calculates temporal feature information, while TV-L1 algorithm is stable in identifying more complex facial motion patterns needed for micro-expression analysis. The method is also improved by the hyperparameter optimization of the method's approach as well as the application of accuracy, Unweighted F1-score, Unweighted Average Recall, Matthews Correlation Coefficient, and Expected Calibration Error as measures for evaluation. Experiments on dataset-- SAMM, SMIC and CASME II databases demonstrate the high accuracy, real-time computation capacity, and capability of the model to learn in dynamic real-world environments. The research bridges the existing gap between current state-of-the-art AI models and real-world emotion recognition systems and offers profound insights into human emotional intricacies. The study highlights the paradigm-reversal value of AI in sensing and responding to human emotions and lays the foundation for even further development of human-focused solutions across many disciplines.**

***Index Terms*— Micro-expression recognition, HTNet, TV-L1 optical flow, facial landmark detection, Expected Calibration Error.**

I. INTRODUCTION

In human-computer interaction and artificial intelligence, understanding human emotions and micro- expressions is essential for the creation of responsive and adaptive systems. Emotions influence decision-making, communication, and perception in daily interactions. By enabling systems to recognize and interpret subtle non-verbal cues, such as micro- expressions, involuntary muscle movements of face that reflect underlying emotions, these systems can facilitate more intuitive and empathetic interactions. This capability is particularly valuable in applications like customer care, interactive media, and personalized medicine, where systems can adapt their responses in relation to the emotional status of a person, ultimately resulting in higher user satisfaction and participation. Micro-expression recognition is also extremely valuable for security and psychology research.

Grenze ID: 01.GIJET.12.2.64_21

In security, it provides a painless way of detecting deception, stress, or discomfort cues, allowing predictive threat assessment and smart decision-making. In psychological research, computerized analysis of fleeting expressions makes large-scale human behavior and affect studies possible, avoiding the constraints of manual collection and interpretation. Moreover, this automation can lead to new solutions in mental health diagnosis and treatment, where understanding emotions is essential for developing personalized treatment plans.

This research utilizes the Hierarchical Transformer Network (HTNet) [1], a deep learning framework specifically chosen for its ability to enhance micro-expression recognition accuracy. Micro-expressions, being fleeting and subtle, require advanced models capable of capturing the intricate details of facial dynamics. HTNet's architecture, which leverages a hierarchical structure of transformer layers, allows it to capture intricate facial dynamics and subtle emotional cues effectively. Preprocessing with the TV-L1 optical flow algorithm is critical in this research as it captures fine facial movements and reduces noise caused by variations in head position and background, ensuring that only relevant features are extracted for classification. By combining HTNet with this preprocessing step, the model can focus on the most informative visual cues, improving both precision and reliability in detecting micro- expressions. The use of TV-L1 optical flow not only enhances the model's focus on significant facial dynamics but also reduces the potential for misclassification caused by irrelevant visual information.

Implementation using benchmark datasets such as Spontaneous Actions and Micro-Movements (SAMM) [2] and Chinese Academy of Sciences Micro-Expression (CASME II) [3] and Spontaneous Micro-expression Database (SMIC) guarantees a valid assessment of the model, and full performance measures guarantee its efficiency. This type of methodology is beneficial to this work in the context that, not only is recognition accuracy enhanced but it also sets a solid foundation for future use in human-computer interaction, security, and psychology research studies where accurate recognition of emotions is of utmost priority. This article contributes to existing advancements in the study of emotion recognition by presenting new solutions that explain human emotions on a fine-grained level.

II. Related Works

This section briefly summarizes recent progress in micro-expression recognition with emphasis on techniques like Attention mechanisms, Capsule Networks, and transformer-based models. The argument points out some major limitations in computational cost, real-time performance, and generalization to datasets, which our proposed solution will overcome.

Attention mechanisms have also garnered widespread interest in micro-expression recognition because they can identify crucial facial areas. M. F. Hashmi et al., [4], introduce LARNet model that utilizes lossless attention residual networks to facilitate recognition enhancement by focusing on the eyes and the mouth, based on spatial and temporal characteristics through residual connections. In these aspects, Li et al. [5] employ spatial & channel attention mechanisms to provide feature information from important facial action units and hence have better recognition rates. These approaches are seen to be extremely robust in terms of precise recognition of small facial movements as well as best preservation of important features.

Capsule Networks provide a new approach to micro-expression recognition through preserving spatial hierarchies. Quang et al. [6] showed that CapsuleNet was able to learn fine-grained spatial relationships effectively, with improved classification performance as a result. Its capacity to preserve spatial relations among features makes it an effective tool in recognition of micro expression.

Automated recognition of micro-expressions has also been investigated for mental health applications. Gilanie et al. [7] proposed a real-time depression detection system by extracting spatial and temporal facial features. The work demonstrates the potential of micro-expression recognition for practical application, e.g., mental health monitoring.

Adegun and Vadapalli [8] further contrasted traditional machine learning classifiers with deep learning approaches for micro-expression recognition. They discovered the dominance of deep models in terms of high accuracy, demonstrating their applicability in multiple datasets.

Cross-dataset recognition is still a significant problem in micro-expression research. This problem was resolved by Jiang et al. [9], who introduced a model that has a greater priority to salient facial areas, improving dataset recognition. The method shows the potential of better generalization and cross-dataset transferability.

Transformers have recently been utilized for micro-expression recognition as they are able to effectively capture temporal dependencies. Nguyen et al. [10] introduced Micron-BERT, a BERT-based model of subtle expression capture with contextual frame-to-frame analysis. Its enhanced performance in detecting temporal patterns makes it a desirable solution in the field.

Li et al. [11] introduced a Deep Local-Holistic Network combining hierarchical convolutional recurrent neural network and robust PCA-based recurrent neural network for spatiotemporal feature extraction and global feature extraction, respectively. Holistic and local features fusion improves benchmark dataset recognition accuracy on CASME, SMIC and SAMM.

Pham et al. [12] presented the Residual Masking Network (RMN) for facial expression recognition. The RMN applies spatial masking to emphasize big facial regions and residual learning to suppress unwanted noises. The method is efficient in detecting normal and subtle expressions.

Lei et al. [13] introduced a new micro-expression recognition system based on facial graph representation learning and facial AU fusion. Using graph-based representation learning and AU information fusion, the model forecasts fine facial movement and achieves outstanding performance on CASMEII, SMIC and SAMM datasets.

The reviewed studies have achieved remarkable success in the creation of micro-expression recognition via techniques such as attention mechanisms, Capsule Networks, transformers, and graph-based learning. Despite their abilities, however, a balance between computation, accuracy, and real-time applicability is still an issue waiting to be fully resolved.

For responding to such demands, we recommend an end-to-end collaborative scheme that fuses TV-L1 optical flow and HTNet. TV-L1 optical flow holds the capacity to capture subtle motion patterns with efficiency and consists of very robustness towards noises and changes in brightness. HTNet is an efficient and up-to-date design which integrates optical flow and transformer-style models for expressing low-computational accurate yet models. The union of the two is highly accurate in identification, computationally efficient, and widely applicable across various datasets and hence ideal to real-world micro-expression recognition tasks.

III. Dataset Description

The system described herein utilizes two pre-existing biometric and behavioral privacy-sensitive datasets Spontaneous Actions and Micro-Movements (SAMM) [2] and Chinese Academy of Sciences Micro-Expression (CASME II) [3]. The work utilizes a combination of the two datasets.

The SAMM dataset includes twenty-eight participants recordings, 132 micro expressions, and 146 long videos with 342 macro expressions. The dataset is also complemented with Action Units coding, including rich data about facial expressions. The dataset provides onset frame, offset frame, and apex frame indicates for micro expressions, providing accurate temporal annotations. The original video samples have resolution 2040× 1088 pixels and 200 frames for a second - frame rate. Emotion classes in SAMM are tagged with labels like "disgust," "fear," "contempt," "anger," "repression," "surprise," "happiness," and "others." There are 91 samples tagged as "negative," 25 tagged as "positive," and 15 tagged as "surprise."

CASME II dataset contains 24 subjects with 144 samples that denote 144 emotions. They were captured in a controlled lab environment where cameras worked at 200 frames per second. Each sample has an original size of 640 × 480 pixels. Emotions are grouped into categories like "happiness," "surprise," "disgust," "sadness," "fear," "repression," and "others." Grouping is into three larger categories, 88 samples with label "negative," 32 with label "positive," and 25 with label "surprise." CASME II also offers accurate onset, offset, and apex indices for every micro-expression.

SMIC-HS dataset contains 16 topics with a total of 164 samples representing 164 emotions. 164 samples are captured through the use of a 100 frames/second lab camera. Size of original samples is 640 × 480 pixels. Facial region of interest is to focus attention on as well as feed consistent input size to our micro-expression recognition task, we resize face images to the size of 28 × 28 pixels. SMIC samples are labeled as "negative", "surprise" & "positive". The count of "negative", "positive", and "surprise" is 70, 51,43. The onset and offset in SMIC, but is not accessible at the apex index in SMIC.

IV. Proposed Methodology

The proposed method reflects a thoroughgoing micro-expression recognition system emphasizing efficient preprocessing, stable feature learning, and accurate classification. In support of recent-state-of-the-art techniques such as the TV-L1 optical flow algorithm to cope with motion dynamics and HTNet for classification relying on deep learning, the system is made vastly accurate and consistent. Upon using the SMIC, SAMM and CASME II datasets, the process has better results in comparison with major challenges on micro-expression recognition relying on powerful assessment and generalization. The following sections describe each component of the pipeline, from data preprocessing to model structure and evaluation.

A. Data Preprocessing

Data preprocessing is one of the most important roles in preparing input data for micro-expression recognition. It involves a sequence of processes, including resizing frames, calculating optical flow, and segmenting the facial region to make sure that the model processes only the right information. Preprocessing begins with resizing the onset and apex frames acquired. Onset and apex frames are the onset and peak of face expressions, respectively. Using the MediaPipe library, such frames are resized so that the dataset becomes homogeneous. MediaPipe ensures uniformity in image size, which is most critical for proper processing by the neural network. Resizing the data makes input data homogeneous, reducing variability caused by different image sizes, resolutions, and aspect ratios of raw datasets. The frames are then downscaled and are used to compute optical flow using the TV-L1 algorithm, one of the key contributions of the presented work.

B. Optical Flow Generation Using TV-L1 Algorithm

The TV-L1 (Total Variation L1) [14] algorithm is utilized in computing the motion between the onset frames and apex frames, registering the micro-expression reflecting small facial movements. TV-L1 is less sensitive than classical optical flow algorithms to noisy data, structure on edges being preserved, and low computational cost.The data fidelity term makes the optical flow very close to pixel intensity differences between frames, and the total variation regularization term constrains smoothness, keeping important edges without over-smoothing important areas.

In this study, TV-L1 computes motion vectors representing pixel direction and magnitude of movement from onset to apex frames. Facial muscle movements, typically non-metrical in static-image analysis, are highlighted by the vectors. Optical flow extraction is central to the detection of dynamic micro-expression components and serves to feed motion-based features to the model for enhanced accuracy in classification. Total variation (TV) regularization with TV-L1 enhances sparsity of the optical flow image and make it computationally efficient with little possibility of overfitting. TV-L1 also isn't very demanding on high light and noise changes sensitivity. TV-L1 also improves the model generalization in general when run on other data like SAMM, CASME II. TV name brings regularization into the fold, which helps in optimization, and it's easier to do for large data. Robustness to partial occlusions (i.e., face parts being occluded or out of view) is another significant robustness. Micro- expressions are eye, mouth, and eyebrow-based movements, and occlusions would occur with head turning or hand movement. TV-L1 performs exceedingly well under these situations and has quality motion vector estimations even in occlusions.

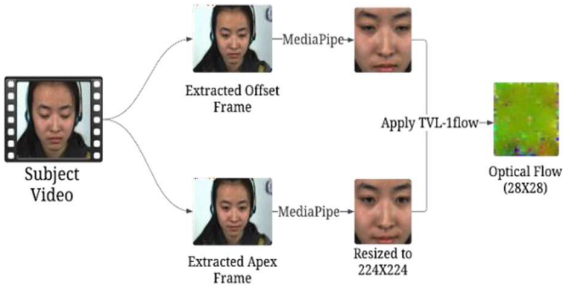

Fig. 1 Optical Flow Generation

C. Facial Region Landmark Coordinates Identification Using MTCNN

The Multi-task Cascaded Convolutional Network (MTCNN)

MTCNN is used for facial landmark detection. [15] is a deep learning model that is formulated to detect the facial regions and key landmark points with high accuracy. MTCNN pre-dicts each frame to provide the coordinates of the eyes, nose, and corners of the mouth. These coordinates are em-ployed to separate the facial area from the rest of the image so that the model will concentrate only on the areas of the face that are responsible for emotional expressions. Align-ing and cropping the facial areas according to these land-marks reduces the differences in head position, orientation, or background noise. This process ensures that what is fed into the neural network is not only clean and uniform but al-so of superior feature quality in later stages. MTCNN's cas-caded, multi-stage structure also allows it to carry out face detection as well as face alignment with extremely high ac-curacy in difficult conditions such as irregular light expo-sure or partial face occlusion. The aggressive preprocessing operation significantly enhances the model's capacity to ac-quire ultra-detailed micro-expression data and guarantees uniform correct recognition performance.

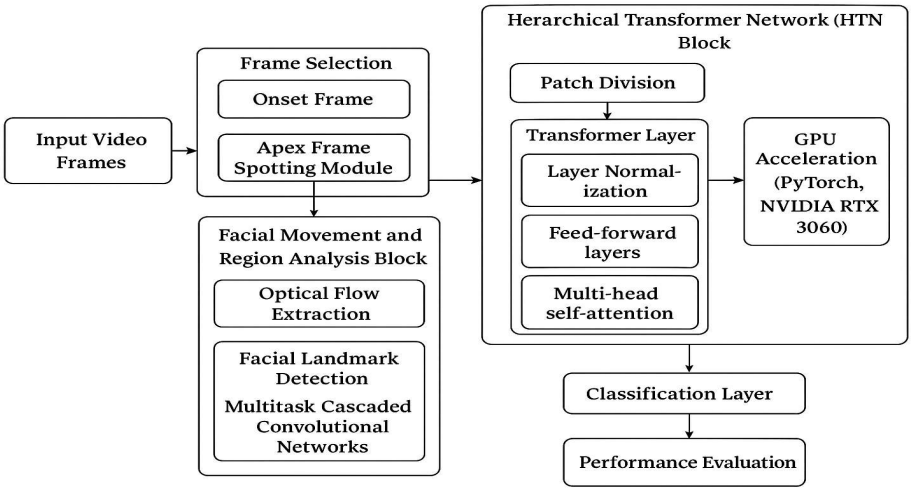

Fig. 2 Architecture Diagram

The HTNet model is a deep learning model specifically designed for micro-expression recognition that can well learn spatial and temporal information. The architectural di-agram is shown in Fig. 2. The pipeline takes pairs of optical flow images, produced by the TV-L1 algorithm, which rec-ord motion dynamics from onset to apex frames. Combined with facial features found by MTCNN, these flow images are used as input to the model, which pays attention to mi-nor facial movements important for classification. Spatial features like wrinkles and muscle movement are found by convolutional layers. Activation and pooling layers are used to reduce computational complexity without sacrificing im-portant details. Long Short-Term Memory (LSTM) layers are used by HTNet to model temporal dynamics of micro-expressions, holding frame dependencies in order to model expression evolution over time. Extracted features are mapped by fully connected layers for classification. Soft-Max output layer yields the predictions for micro-expression classes, while Expected Calibration Error (ECE) provides reliability in predictions.TV-L1 output as sparse motion vectors can conveniently be incorporated within HTNet since these motion vectors are meaningful representations to be subsequently processed with recurrent or convolutional layer

V. Experimental Analysis and Results

Evaluation was performed on CASME II, SMIC, SAMM, and Full datasets using preprocessing methods such as onset and apex frame separation, resizing using MediaPipe, and motion data extraction using TV-L1 optical flow algorithm for the purpose of detection of subtle facial dynamics essential for micro-expression analysis. Training was performed on a computer equipped with an NVIDIA GeForce RTX 4050 GPU to facilitate efficient processing of high-dimensional input and iterative hyperparameter tuning. Training and inference were performed on an NVIDIA GeForce RTX 4050 GPU with CUDA 12.1. GPU acceleration significantly reduced training time and enabled real-time processing of optical flow vectors. Although TV-L1 would be using more computation resources, the amount of computation can be handled with a high-speed GPU like NVIDIA, GeForce RTX 4050. Use of the GPU assists us with the computational need without decreasing the real time-usability.

The HTNet model had a hierarchical architecture with the input image of size 28 and patch of size 7 projected to the embedding dimension of size 256. The model had three hierarchical levels, with transformer block configurations set as (2, 2, 10). Hyperparameters, such as embedding dimensions and the number of attention heads, were fine-tuned based on the observed trends in Unweighted F1-score (UF1) vs Dimensions, Unweighted Average Recall (UAR) vs Dimensions, UF1 vs Heads, and UAR vs Heads. Specifically, UF1 vs Dimensions and UAR vs Dimensions guided the selection of the embedding dimension by identifying the optimal balance of performance across datasets at various dimensional configurations. Similarly, UF1 vs Heads and UAR vs Heads helped determine the adequate number of attention heads to best capture fine-grained feature interactions without overfitting. These analyses allowed the model to best classify micro-expressions into three groups: positive, negative, and surprise. Hyperparameter tuning also uncovered dataset-specific sensitivities, showing the resilience of CASME II under conditions of increased dimensionality as well as different numbers of attention

heads, while SAMM and Full datasets showed the necessity for more conservative optimization. With these performance trends, HTNet was optimized to be maximally accurate and robust so that it can detect fine-grained micro-expressions effectively on a broad range of datasets.

Where class distributions are imbalanced, in cases where the class distribution of a data set includes surprise, positive, and negative with a 1:1.31:3 ratio, performance measures like the Unweighted F1-score and Unweighted Average Recall reflect an even perspective of classification accuracy. The UF1 measure is especially valuable in multi-class settings with imbalanced data sets because it computes F1-scores for all classes without comparing them to their frequencies.

The class-specific F1-score $F1_c$ is computed as:

$$F1_c = \frac{2 \cdot TP_c}{2 \cdot TP_c + FP_c + FN_c} \quad (1)$$

where TP_c, FP_c, and FN_cdenote true positives, false positives, and false negatives for class c, respectively. It represents the harmonic mean of precision and recall, effectively balancing both to measure classification performance.

UAR guarantees the model's recall performance is tested for every class irrespective of class size, also ensures the model's capability to accurately classify all sample classes and remove bias towards the majority Fclass and its calculated as:

$$UAR = \frac{1}{C}\sum_{c=1}^{C} \frac{TP_c}{TP_c + FN_c} \quad (2)$$

A. UF1 VS Dimensions

The HTNet model performance was tested on Full and all three dataset i.e SMIC, SAMM, and CASME II datasets using UF1 measure for dimensionalities 112, 192, 256, and 392. CASME II performed best in all scenarios with an all-time best UF1 score of 0.9722 at 256 dimensions but fell to 0.86 at 512 dimensions due to overfitting or too much feature noise. SMIC rose to 256 dimensions, a peak of 0.788, before falling to 0.68 at 392 dimensions, indicating issues with high feature space. Full dataset exhibited consistent performance, with a peak of 0.864 at 256 dimensions, but falling to 0.825 at 392 dimensions because of larger variability from dataset combi-nation. The results highlight 256 dimensions as best for achieving feature representation and model performance bal-ance. Consistent performance of CASME II indicates its suit-ability for micro-expression analysis, and the Full dataset implies variability issues. Dimension reduction could be in-vestigated in future research to reduce losses at higher di-mensions and make models more consistent.

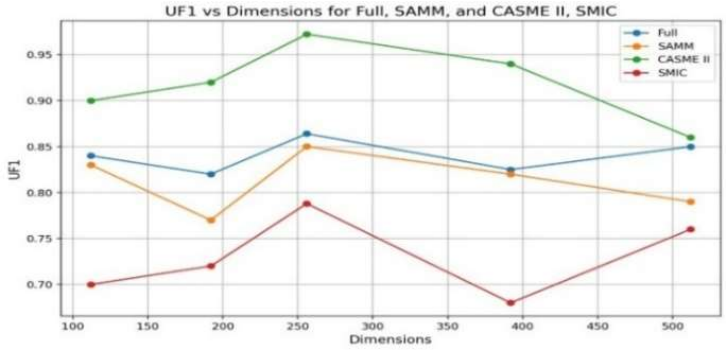

Fig.3 UF1 vs Dimensions

B. UAR VS Dimensions

The performance is measured for the UAR (Unweighted Average Recall) on Full, SMIC, SAMM, and CASME II datasets for various dimensionalities. CASME II always provides the best performance with the highest UAR of 0.9658 at 256 dimensions, and SMIC with the highest improving to 0.7929 at 256 dimensions. The Full dataset provides consistent performances with highest UAR of 0.8701 at 256 dimensions. The outcome indicates that CASME II takes advantage of greater dimensions, where its strength is revealed. SAMM, although initially weaker, is stronger with greater dimensions, i.e., sensitive to richness in features. Full dataset shows more stable performance, which reveals its flexibility but low sensitivity to high-dimensional features. These outcomes highlight the significance of optimal dimensions, especially in datasets such as SMIC, CASME II and SAMM.

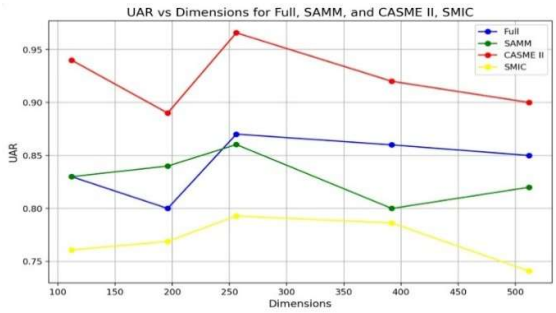

Fig.4 UAR vs Dimensions

C. UF1 VS Heads

The study analyzes the impact of attention heads on UF1 scores for the Full, SMIC, SAMM, and CASME II datasets. CASME II is consistently better with the best UF1 score of 0.9722 at 3 heads and remains consistent across settings. The Full dataset has an optimum of 0.864 at 3 heads but deteriorates with increments thereafter. SMIC also peaks at 3 heads with a UF1 of 0.788 before performance declines. The findings highlight the need to choose the best number of attention heads, with all the datasets achieving the best performance at 3 heads. CASME II is most robust even in suboptimal conditions, whereas SAMM, SMIC and Full datasets are sensitive and decline more precipitously beyond the best setting. This draws attention to balancing between model performance and model complexity across datasets.

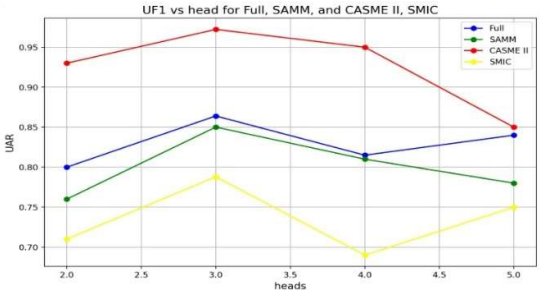

Fig.5 UF1 vs Heads

D. UAR VS Heads

The evaluation contrasts UAR performance of UAR for the Full, SMIC, SAMM, and CASME II datasets under different attention head settings. CASME II has the best UAR throughout, with 0.9658 when it is set at 3 heads and remaining stable. The Full dataset shows significant improvement, with its peak at 3 heads at 0.8701 but falling dramatically at higher heads. SMIC is the same, with its peak at peak of 0.7929 at 3 heads after which it declines by a significant margin. The output indicates that CASME II is less sensitive to the change in number of heads and demonstrates stable performance across settings. Full, SMIC and SAMM datasets, however, are best at 3 heads, after which their performance declines, and hence optimal hyperparameter tuning for these datasets is needed in order to display best performance.

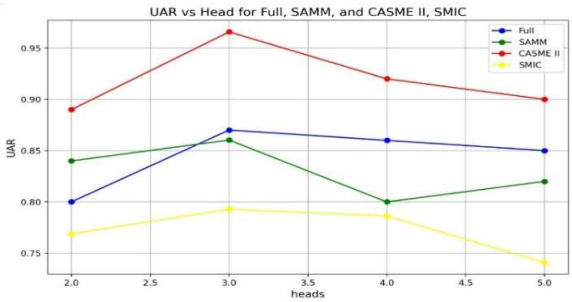

Fig.6 UAR vs Heads

E. Other Metrics Used

Apart from UF1 measurement, the performance of model was also checked by Matthews Correlation Coefficient (MCC) and Expected Calibration Error (ECE) to encompass more comprehensive evaluation. MCC, being a good estimator, measures quality of classification based on all the components of confusion matrix positives true (TP), true negatives (TN), false positives (FP), and false negatives (FN). It is computed using the following formula:

$$MCC = \frac{(TP \cdot TN) - (FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (3)$$

During this experiment, Weighted Average MCC was 0.6361, which reflected moderate performance with respect to class imbalances. Macro-Averaged MCC, with equal weightage to all classes, was slightly lower at 0.5352, indicating difficulty in minority class classification. Overall MCC, a cumulative measure of quality in classification, was 0.8082, reflecting good reliability in classification. Expected Calibration Error quantifies the agreement between predicted probabilities and actual outcomes, giving an idea of the level of confidence in the model. Obtained ECE combining all 3 dataset is 0.1131. ECE is calculated as:

$$ECE = \sum_{m=1}^{M} \frac{|B_m|}{n} \cdot |\mathrm{ACC}(B_m) - \mathrm{CONF}(B_m)| \quad (4)$$

where $B_m$is the m^{th}confidence bin, n is the total number of samples, $\mathrm{ACC}(B_m)$is the accuracy in that bin, and $\mathrm{CONF}(B_m)$is the average predicted confidence. It measures the calibration gap between predicted probabilities and actual accuracies.

VI. CONCLUSION

In this paper, the Hierarchical transformer Network, one of the most recent micro-expression recognition frameworks, takes advantage of the hierarchical nature of transformer networks and the superior preprocessing features of the TV-L1 optical flow algorithm. The TV-L1 outperforms other approaches to record the fine facial dynamics through edge detail preservation and noise suppression, and specifically very efficient in identifying the subtle facial movements involved in micro-expressions. With its ability to get motion vectors from apex to onset frames, it is possible to represent temporal dynamics accurately even on noisy or poor-quality datasets. HTNet does not only pose new opportunities to micro-expression identification but also achieves a standard to adaptive emotion perception systems with the novel application of TV-L1 optical flow as well as transformer-based architectures. HTNet with TV-L1 optical flow has state-of-the-art performance on the SMIC, SAMM and CASME II databases, solving issues such as varied micro-expression categories and data class imbalance. Its utilization of sophisticated performance metrics, including MCC and ECE, further supports the validity of the framework. With its applications including human-computer interaction, psychological analysis, and security, HT Network is concerned with its applicability to real-world deployment.

ACKNOWLEDGMENT

The authors wish to thank JSS Science and Technology University and JSS Mahavidyapeeta.

REFERENCES

[1] Wang, Zhifeng & Zhang, Kaihao & Luo, Wenhan & Sankaranarayana, Ramesh. (2023). HTNet for micro-expression recognition. 10.48550/arXiv.2307.14637.

[2] A. K. Davison, C. Lansley, N. Costen, K. Tan and M. H. Yap, "SAMM: A Spontaneous Micro-Facial Movement Dataset," in IEEE Transactions on Affective Computing, vol. 9, no. 1, pp. 116-129, 1 Jan.-March 2018, doi: 10.1109/TAFFC.2016.2573832.

[3] Yan, Wen-Jing & Li, Xiaobai & Wang, Su-Jing & Zhao, Guoying & Liu, Yong-Jin & Chen, Yu-Hsin & Fu, Xiaolan. (2014). CASME II: An Improved Spontaneous Micro-Expression Database and the Baseline Evaluation. PloS one. 9. e86041. 10.1371/journal.pone.0086041

[4] M. F. Hashmi, B. K. Kumar Ashish, V. Sharma, A. G. Keskar, N. D. B. Bokde, J. H. Yoon, and Z. W. Geem, "LARNet: Real-time detection of facial micro expression using lossless attention residual network," Sensors, vol. 21, no. 4, p. 1098, Apr. 2021.

[5] Y. Li, X. Huang, and G. Zhao, "Micro-expression action unit detection with spatial and channel attention," Neurocomputing, vol. 436, pp. 221- 231, 2021.

[6] N. V. Quang, J. Chun and T. Tokuyama, "CapsuleNet for Micro- Expression Recognition," 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019), Lille, France, 2019.

[7] G. Gilanie, M. ul Hassan, M. Asghar, A. M. Qamar, H. Ullah, R. U. Khan, N. Aslam, and I. U. Khan, "An automated and real-time approach of depression detection from facial micro-expressions," Computers, Materials & Continua, vol. 73, no. 2, pp. 2513-2528, 2022.
[8] I. P. Adegun and H. B. Vadapalli, "Facial micro-expression recognition: A machine learning approach," Scientific African, vol. 8, p. e00465, 2020.
[9] X. Jiang, Y. Zong, W. Zheng, J. Liu, and M. Wei, "Seeking salient facial regions for cross-database micro-expression recognition," in 2022 26th International Conference on Pattern Recognition (ICPR), Montreal, QC, Canada, 2022, pp. 1019-1025.
[10] Nguyen, Xuan-Bac, Chi Nhan Duong, Xin Li, Susan Gauch, Han-Seok Seo, and Khoa Luu. "Micron-bert: Bert-based facial micro-expression recognition."Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1482-1492. 2023.
[11] J. Li, T. Wang, and S.-J. Wang, "Facial micro-expression recognition based on deep local-holistic network," Applied Sciences, vol. 12, p. 4643, 2022.
[12] L. Pham, T. H. Vu and T. A. Tran, "Facial Expression Recognition Using Residual Masking Network," 2020 25th International Conference on Pattern Recognition (ICPR), Milan, Italy, 2021
[13] L. Lei, T. Chen, S. Li and J. Li, "Micro-expression Recognition Based on Facial Graph Representation Learning and Facial Action Unit Fusion," 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Nashville, TN, USA, 2021, pp. 1571- 1580.
[14] Zach, C., Pock, T., & Bischof, H. (2007). A duality-based approach for realtime TV-L1 optical flow. In Proceedings of the 29th DAGM Symposium on Pattern Recognition (pp. 214-223). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-74936-3_22
[15] K. Zhang, Z. Zhang, Z. Li and Y. Qiao, "Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks," in IEEE Signal Processing Letters, vol. 23, no. 10, pp. 1499-1503, Oct. 2016, doi: 10.1109/LSP.2016.2603342.