Hybrid Real Time Facial Emotions Recognition on Autistic Individuals

Fatima Ezzahrae El Rhatassi[1], Btihal El Ghali[2], Najima Daoudi[3]
ITQAN Team-LYRICA Lab, Information Sciences School (ESI), Rabat, Morocco[1, 2]
SSLab-ENSIAS, Mohammed V University, Rabat, Morocco[3]

Abstract—**Communication and social interaction issues are frequently linked to autism, which can have an impact on quality of life, work, and education. Opportunities to lessen these difficulties are presented by assistive technology, especially those that facilitate individualized and encouraging engagement. Facial expression recognition (FER) is essential to these systems, but current methods are still inadequate for autism-specific situations even though they achieve high accuracy on benchmark datasets like CK+. Because autistic people usually exhibit aberrant, subtle, or ambiguous facial expressions—which deviate from the common patterns used to train traditional FER models—this limitation occurs. In this work, we suggest a hybrid model that combines an LSTM network for temporal modeling of video sequences with three pretrained convolutional neural networks (EfficientNetB0, ResNet50, and MobileNetV2) for spatial feature extraction. Although the model performs well on CK+, its applicability to autism is still limited by the lack of relevant datasets and the use of artificial intelligence (AI)-generated videos rather than authentic recordings. The critical need for more comprehensive data and adaptive model designs tailored to autistic populations is highlighted by these findings.**

Keywords—Facial emotion recognition; video; frames; spatial and temporal features; pretrained models; LSTM; autistic individuals

I. Introduction

Each of us needs to communicate since we belong to a community. Some people with autism spectrum disorders experience communication challenges. These individuals struggle to acquire language skills such as speaking and expressing their thoughts and emotions [1]. Emotion processing plays a critical role to help the autistic individuals face those challenges [2]. For this reason, we considered concentrating on the emotions of autistic people, which can be examined and identified from an image or live video when they communicate with a chatbot designed specifically for this population. While existing video-based FER models achieve good performance on standard datasets, they often fail to capture the atypical and subtle expressions of autistic individuals, due to muted intensity, ambiguous variations, and non-standard temporal dynamics. The principle behind this project is that we try to use a chatbot that provides to the autistic a personalized healthcare through conversation exchange. The chatbot will analyze the autistic person's words and attempt to identify his facial emotions and sentiments in order to help him feel better and avoid any urgent situations wherein he might need to notify parents, medical assistance, or other parties. Accordingly, the main objective of this study is to develop, evaluate, and adapt a hybrid FER model that combines multiple pretrained CNNs (EfficientNetB0, ResNet50, and MobileNetV2) with an LSTM for temporal modeling. The research is guided by the following question: Can such a multi-feature CNN–LSTM architecture improve the recognition of atypical autistic facial expressions in real-time video sequences? In this work, we limit the scope to facial expressions only, without including posture, gestures, or eye contact. While autistic people's emotions have been detected by picture analysis in earlier work [3], in this study, we try to uncover the challenges that arise while analyzing an autistic person's face in real time.

To maintain the logical structure, we will discuss the following areas: The related works that used video-based facial expression identification techniques rather than image-based ones will be shown in Section II. Our proposed work will be presented in Section III and the findings will be addressed. In Section IV, we discuss the usefulness and significance that this research has for individuals with autism. Section V presents the conclusion and future work.

II. Related Work

The majority of existing efforts have focused on the independent analysis of static images and their extension to videos, aiming to identify a person's emotion based on facial cues in those frames [4]. Consequently, research has shifted from relying solely on spatial feature extraction toward incorporating temporal feature modeling to capture the dynamic evolution of expressions. In parallel, a few initiatives have attempted to address real-time facial emotion recognition (FER) specifically for individuals with autism. For instance, Talaat et al. [22] developed a CNN and IoT-based framework, achieving up to 94.6% accuracy in real-time conditions while reducing latency through for computing. Similarly, Abou El-Magd et al. [23] proposed a hybrid kernel autoencoder–CNN model, which obtained 96.7% training accuracy and 95.4% validation accuracy on small autism-related datasets. Other efforts focused more on assistive systems: Lee and Wong's AEGIS [24], an augmented reality application, reported real-time emotion recognition with an average accuracy of 92%. However, such studies remain scarce and are often constrained by small or synthetic datasets, limited reporting of latency under real-world conditions, and the absence of large-scale evaluations on naturalistic autistic populations. Thus, while promising, current autism-focused FER research still lacks the robustness and generalizability of benchmark-oriented studies (e.g., CK+), which underscores the need for spatio-temporal models optimized for real-time use and trained on representative autistic datasets. For example, Jad Haddad and his associates [5] think that temporal information is essential for tracking minor facial changes that occur when an emotion is expressed. They tried to build an efficient 3D-CNN for emotion recognition in videos, which takes as input a collection of sequential frames and analyzes them collectively, while maintaining the temporal aspect of a video sequence. The Optuna framework for effective hyper-parameter search was utilized to identify the best architecture for the 3D CNN model, which also contained the temporal hyperparameters. According to certain 2017 studies [6], the mouth and eye are the most crucial features for recognizing facial emotions. The eye and mouth areas are extracted using Temporal Gabor Features. Following normalization using the Z-score normalization procedure, they are encoded into binary pattern features, which are then concatenated to produce improved temporal features. Performance on the RML and CK datasets is significantly improved by this technique.

Some people use HappyER-DDF, a technique designed to identify pleasant emotions in videos [12]. In order to extract spatial-temporal data from sequential video frames, this system used a 3D-Inception-ResNet neural network. This method is novel by the fact that it incorporates a Long Short-Term Memory (LSTM) unit to record temporal dynamics, which is essential for precisely identifying emotions in videos.

Several CNN models, including VGG-Face, ResNet18, Densenet121, and VGG16, are used in other publications as [7] to handle facial images. A two-layer Bi-LSTM is then used to collect dynamic information. By combining each model's advantages, the fusion approaches seek to increase overall accuracy [8], [9], [10]. It can involve applying Principal Component Analysis (PCA) to a single emotional video in the two phases of preprocessing and feature extraction, combining other CNN architectures in the emotion classification phase [11], and analyzing emotion expressions using models such as VGG-FACE, LSTM (Long Short-Term Memory networks), and Xception with a statistic encoder (STAT) [9]. Generally the use of models or the idea of combining many models or architectures depend on the computational and time constraints in addition to the dataset used on training that impact automatically the results.

Facial expression recognition has advanced significantly in recent years, according to numerous research studies. For spatiotemporal feature extraction, some suggest merging deep learning and dynamic texture techniques, such as LSTM and VGG19 [13]. By incorporating a conventional Vision transformer (VIT), multi-View Complementary prompters (MCPs), and temporal-Modeling Adapters (TMAs) into the image model, others attempted to modify static image models for dynamic video-based emotion recognition [14]. In addition to face emotions, physiological information such as heart rate are presented on a video-based multimodal [15]. To racalibrate the channel-wise feature maps, a new framework is proposed by combining a 3D Xception network with Squeed-and-excitation modules. The HRV characteristics and iPPG signals were chosen as the two parameters that best illustrate the feasibility and promise of multimodal fusion.

There aren't many datasets utilized on this subject in general; we can include AFEW, CK+, and MMI, and they typically require a demand from their owner in order to be used. By focusing on temporal features and employing a variety of methods, including handcrafted descriptors, spatio-temporal aggregation, multimodal fusion, and pre-trained models, each paper advances video-based face emotion recognition. The various methods employed in various studies for facial expression identification, the dataset utilized, and the model's accuracy have been gathered in the next Table I.

TABLE I. Papers Contributions and Techniques to Advancing Video-Based Facial Emotion Recognition

Objectives	Dataset	Techniques	Accuracy
Enhance emotion recognition in visual data by employing a deep neural network-based fusion model [9]	AFEW 2016, SAVEE, CEV data	VGG-FACE LSTM Xception	Accuracy of 54.83%
Recognizing happy emotions from videos using combined 3D visual information with deep learning techniques [6]	AM-FED+, AFEW, MELD	3D-Inception-ResNet	Accuracy of 95.97%
Recognize the Facial Emotion in Video Sequences Using Eye and Mouth Temporal Gabor Features [12]	RML and CK datasets	Gabor Wavelets, ensemble classifier, temporal feature encoding	The accuracy of Temporal feature extraction exceeds 90% for most of the emotions
Bi-modality Fusion for Emotion Recognition in the Wild [17]	AffectNet dataset, EmotiW	Video: VGG16,CNN,Bi-LSTM Audio: low level descriptor (LLDs)	Accuracy of 62.78%
EmotiEffNet and Temporal Convolutional Networks in Video-based Facial Expression Recognition and Action Unit Detection [13]	eNTERFACE05 database	VGG19, LSTM, HOG-HOF descriptors, Support Vector Machine (SVM)	Accuracy of 98%
From Static to Dynamic: Adapting Landmark-Aware Image Models for Facial Expression Recognition in Videos [14]	RAF-DB, AffectNet7/8 , and FERPlus, DFEW, FERV39K, and MAFW	Landmark-aware pre-trained models, temporal adapters	Accuracy varies from 25% to 93%. It depends on the emotion.
Video-Based Multimodal Spontaneous Emotion Recognition Using Facial Expressions and Physiological Signals [15]	BP4D+	Multimodal fusion of facial and physiological features	Accuracy of 70%
Real-Time Facial Emotion Recognition Using Deep Learning Models [16]	BAUM-1s, eNTERFACE05	3D Convolutional Neural Networks, spatial–temporal feature aggregation	Around 45% of accuracy

III. Proposed Work: The Fusion of Three Models

The identification of facial expressions from images or videos has been the subject of numerous studies; nonetheless, the fundamental framework is essentially composed of the same components as shown in Fig. 1.

At the data collection stage, the focus is on gathering facial expression datasets from videos. Several datasets are publicly available, and among them, we chose CK+. This choice was motivated not only by the fact that CK+ has been pretrained on a large amount of data, but also because it was the only dataset provider who responded positively to our request and granted us access. In addition, CK+ presents significant advantages compared to static image datasets, as it contains complete temporal sequences:

- Neutral starting point: Each sequence begins with a neutral facial expression.
- Progressive evolution: The dataset captures the gradual development of the expression until its peak intensity.
- Detailed annotations: CK+ provides more than just a global label for each video.
  - Sequence-level labels: Each video sequence is labeled with a single primary emotion, ensuring clarity for training.
  - Apex frame annotations: The exact frame where the expression reaches its peak intensity is explicitly identified.

These characteristics make CK+ particularly suitable for our study, especially when combined with a fusion of pretrained models—EfficientNetB0, ResNet50, and MobileNetV2—used for the training stage.

Generally, the preprocessing stage involves identifying faces in every video frame utilizing deep learning-based or face identification algorithms [17]. We included cropping and aligning faces to guarantee uniformity in facial areas between frames and adjusting the photos for changes in scale, position, and lighting. It is essential to gather temporal and spatial feature [18, 19] analysis in order to build a performant model that considers the relationship between the frames in order to recognize face emotions from a video.

The architecture that we are proposing is illustrated below in Fig. 2.

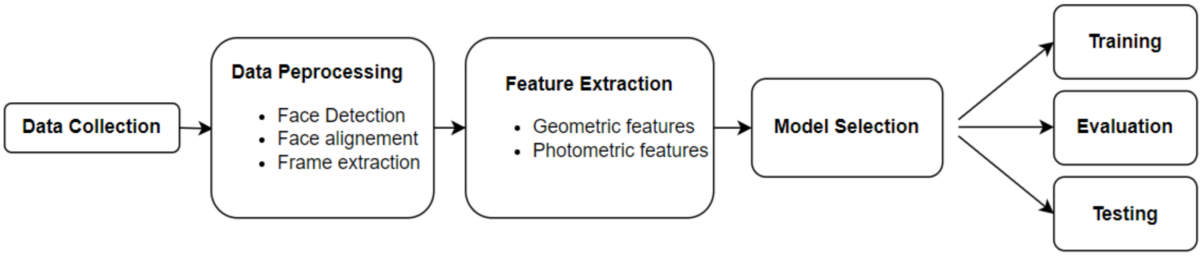


Fig. 1. The process of facial emotion recognition.

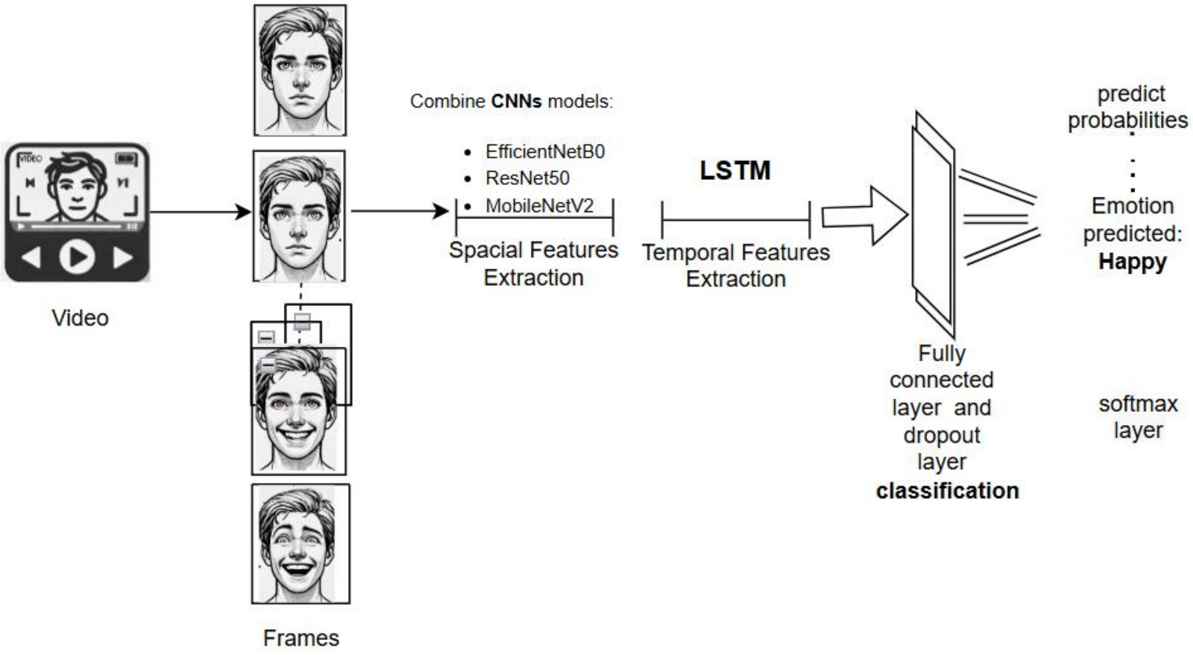


Fig. 2. The model of facial emotion recognition by combining CNNs models and LSTM for features extraction.

We will be using the combination of CNN (Convolutional Neural Networks) for spatial feature extraction and RNNs (Recurrent Neural Networks) for temporal dependencies by the use of LSTM (Long Short-Term Memory).

In order for our model to learn from the CK+ dataset, which is made up of folders with sequential frames but no individual image labels, the first step is to label these sequential frames. Each folder in the CK+ dataset represents an emotion sequence, with the final frames showing the peak of an emotion (such as happy, angry, or surprised) and the initial few frames usually showing a neutral face. In order to read sequences (video frames) and the labels that go with them, we first load the CK+ dataset (pictures and labels) into memory for the training and validation phases. We then cycle through image folders. Then, for multi-class classification, we use to_categorical to transform labels into one-hot encoding.

For efficiency, we preprocess the data before storing it in npy files and utilizing train_test_split (80%-20% split) to divide the dataset into training and validation sets.

We create a Multi-Feature Base Model by creating a base feature extractor model using the pre-trained networks EfficientNetB0, ResNet50, and MobileNetV2. We extract features from three pre-trained models for each frame and use TimeDistributed to apply the same feature extractor across all frames in a sequence and finally combine features from all three models using Concatenate. After that we add a LSTM layer to model temporal dependencies in the feature sequences and for the classification we add fully connected and dropout layers and final output layers uses softmax activation to predict probabilities for eight classes (anger, contempt, disgust, fear, happiness, neutrality, sadness and surprise).

The choice of this method is motivated by several factors:

- Complementary strengths of CNN and LSTM: CNNs excel at capturing spatial details from facial regions, while LSTMs are effective in modeling the temporal evolution of expressions across frames.
- Robustness through model fusion: By combining EfficientNetB0, ResNet50, and MobileNetV2, we leverage the diversity of their learned representations, improving generalization compared to using a single model.
- Efficient handling of temporal sequences: The use of TimeDistributed allows frame-by-frame processing while maintaining temporal coherence, which is crucial for datasets like CK+ that provide full expression sequences.
- Scalability and flexibility: The architecture can be easily extended to incorporate additional pretrained models or adapted to other datasets with minimal adjustments.
- Improved classification accuracy: The multi-feature extraction, combined with LSTM's temporal modeling, enhances the system's ability to capture subtle expression changes, which are often missed in static-image approaches.

Following these processes, it's time to assess the model, examine how well it performs across a range of emotions, and determine its limitations.

IV. Experimental Results

A. Applying the Model to CK+ Dataset

Fig. 3 illustrates the training and validation accuracy and loss curves over the course of the training process, providing insights into the model's performance and generalization.

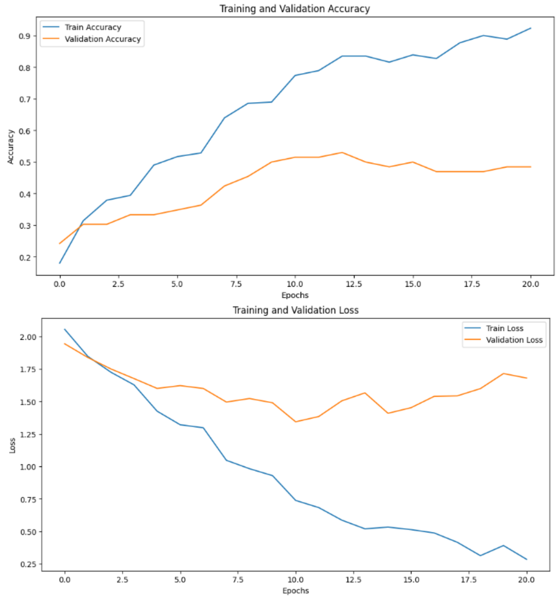


Fig. 3. The model's training and validation accuracy/loss.

1) Accuracy and Loss: Between Epochs 1 and 21, the training accuracy increases progressively from 17% to 94.5%. The model is successfully learning from the training data when the training loss steadily drops.

2) Validation Accuracy and Loss: In Epoch 13, the validation accuracy peaked at 53.03%, having begun at 24.2%. After that, though, it starts to plateau and even slightly declines (for example, 48.48% in Epoch 21). After initially declining, the validity loss starts to change at about Epoch 10. This suggests that the model has trouble generalizing outside of the training set.

3) The model's correct predictions: Here in Fig. 4 are some of correctly classified images along with their corresponding true labels and predicted labels that visually inspects how well the model has performed on correctly predicted samples.

4) The model's false predictions: The model successfully learns from the training data but struggles to generalize, as evidenced by the images below in validation data in Fig. 5.

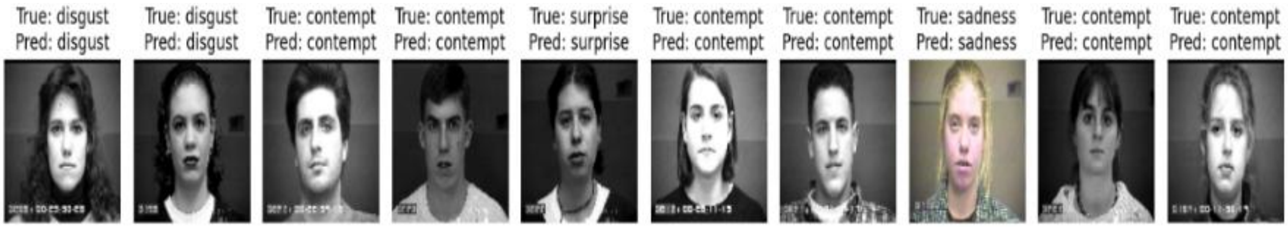


Fig. 4. A sample of correctly predicted pictures.

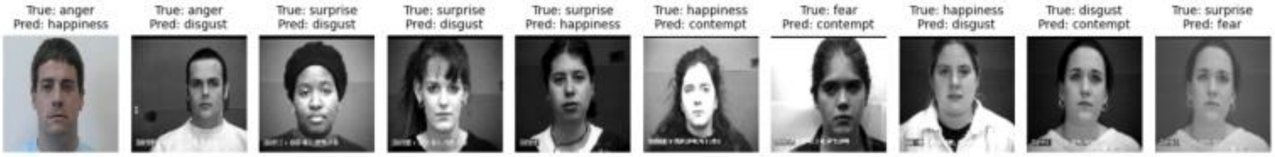


Fig. 5. A sample of incorrect predicted pictures.

In our experiments, we implemented a hybrid architecture combining CNNs and LSTMs. Spatial features were extracted using three pretrained backbones (EfficientNetB0, ResNet50, and MobileNetV2), whose outputs were concatenated to form a multi-feature representation. Temporal dependencies across video sequences were then modeled using LSTM layers, followed by fully connected layers for classification.

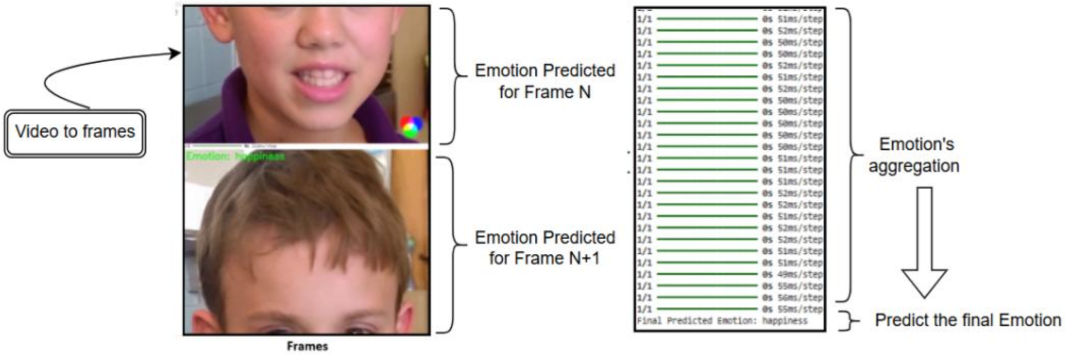
While the training accuracy reached 94.5 % after 21 epochs, the validation accuracy plateaued around 53 %. This performance is significantly below the state-of-the-art that worked with the same dataset as [20,21], where recent works report accuracies above 95–98 % on CK+ using attention-based CNNs or spatio-temporal models such as FAN or CAGNet. We attribute this gap mainly to overfitting due to the high model complexity relative to the dataset size, and the absence of attention mechanisms to focus on the most informative frames or facial regions.

Nevertheless, our architecture demonstrates the feasibility of combining multiple pretrained models with temporal modeling, aligning with current research trends. Future work will focus on reducing overfitting through data augmentation, regularization, and possibly integrating attention modules to improve generalization and bring performance closer to the reported state-of-the-art.

B. Applying the Model to Videos of Autistic Children

In our last research [2], we worked with images of autistic children available online, gathered by Fatima Talaat at https://www.kaggle.com/datasets/fatmamtalaat/autistic-children-emotions-dr-fatma-m-talaat. Since we face a huge problem related to data and privacy, especially when it comes to people with autism, it is so hard to get acceptance to use the data even if it will not be shared. The lack of available datasets that can respond to this need in order to validate our proposed approach made us think about using the artificial intelligence tools to make videos only from the images that we worked with previously. The tool used in our case is https://www.HeyGen.com/ which is an AI-driven platform that enables users to create AI-generated videos quickly and efficiently. By selecting an image of the autistic child and expressing the setting and emotion we wanted it to convey, we attempted to produce a customized video content. We then applied our fusion model to it.

The steps to predict the final emotion are as follows in Fig. 6.


Video to frames
Emotion: happiness
Frames
Emotion Predicted for Frame N
Emotion Predicted for Frame N+1
Emotion's aggregation
Predict the final Emotion

```
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 51ms/step
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 52ms/step
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 50ms/step
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 50ms/step
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 52ms/step
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 51ms/step
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 52ms/step
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 50ms/step
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 50ms/step
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 50ms/step
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 50ms/step
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 50ms/step
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 51ms/step
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 51ms/step
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 51ms/step
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 52ms/step
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 52ms/step
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 51ms/step
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 52ms/step
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 52ms/step
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 51ms/step
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 51ms/step
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 49ms/step
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 55ms/step
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 56ms/step
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 55ms/step
Final Predicted Emotion: happiness
```

Fig. 6. The schema of splitting frames and predicting the final emotion.

This schema describes the pipeline for video-based emotion recognition, focusing on how frames are processed and predictions are made for sequences of frames in our model application.

1) Video processing by breaking the video containing the autistic into frames: A video is essentially a sequence of static images (frames) displayed over time. The first step is to extract these individual frames from the video and to preprocess only frames with faces.

Each frame is treated as an image, enabling the model to process it independently or in groups. Frame-by-Frame Processing: This method ensures that every segment of the video is examined chronologically by processing it frame by frame. Since emotions in a video frequently change over time, this is essential for preserving the temporal context.

2) Sequence-based emotion prediction by processing sequences of frames: The pipeline aggregates frames into sequences of a fixed size (e.g. time_steps = 10) rather than predicting emotions for a single frame separately. This method captures temporal dependencies, which are crucial for comprehending dynamic emotions because of the interaction between successive frames.

3) Emotion's prediction where the model predicts the corresponding emotion (e.g., happy, sad, angry...) for every set of frames. These sequences are handled by temporal models, LSTMs which can learn from patterns and changes in the frames over time.

4) Aggregation of predictions by applying dominant emotion inference: The model combines the predictions made after analyzing every frame sequence to identify the main emotion expressed in the video. For example, if the majority of the scenes are categorized as "happy," then "happiness" may be the general feeling conveyed by the video. Voting, averaging probability, or choosing the most frequent forecast are some examples of aggregation techniques.

5) Result of applying the model on videos containing autistic individuals: The model performs well overall but struggles with many false predictions. These can be related to some autistic children who exhibit atypical or subtle facial expressions of emotions as muted smiles or ambiguous anger or fear expressions, which differ from what we had in the training data in terms of the expressed facial emotions. Also, the generative AI videos may not perfectly capture real-world variations or traits in autistic expressions, which can help create bias.

Our model needs some improvements in having more diverse, real-world dataset of autistic expressions and architecture's adjustment to focus on subtle facial regions as mouth corners, eyebrows.

Several challenges arise when applying FER in real time for individuals on the autism spectrum:

- Subtlety and variability of expressions: Autistic individuals may display emotions in less pronounced or non-standardized ways, making it difficult for models trained on typical datasets to recognize them accurately.
- Reduced or inconsistent eye contact: Since many FER systems rely heavily on eye-region cues, this can lead to incomplete or misleading feature extraction.
- Contextual dependence: Some expressions may not be directly tied to a single emotion but are influenced by the situational context, requiring the system to interpret beyond facial features.
- Sensory and behavioral differences: Atypical movements, self-stimulatory behaviors (stimming), or avoidance of camera focus may interfere with real-time face detection and tracking.
- Latency and real-time constraints: Achieving both high accuracy and low processing delay is particularly challenging when expressions are subtle and rapidly changing.
- Dataset limitations: The scarcity of large, diverse, and ethically collected datasets of autistic individuals further restricts the model's ability to generalize.

To address these issues, our model requires improvements in two main directions: 1) access to more diverse and representative real-world datasets of autistic expressions, and 2) architectural adjustments that give greater attention to subtle facial regions such as mouth corners, eyebrow movements, and micro-expressions.

V. Conclusion and Future Work

For the identification of emotions in videos, our approach is based on an architecture that allows the combination of several pre-trained models. We chose to merge the three pre-trained CNNs (EfficientNetB0, ResNet50, and MobileNetV2) with a temporal model based on LSTM in order to extract both spatial features (using CNN) and temporal relationships (using LSTM) from the video sequence. On the CK+ dataset, our model showed strong learning capability, with training accuracy reaching 94.5%. However, the validation accuracy plateaued at 53.03%, which highlights the model's limited ability to generalize beyond the training data and confirms the difficulty of applying FER to real-world autistic expressions. Because we intend to interact with a group of people with autism, and since their facial expressions in a variety of scenarios are unique, recognizing emotion is difficult even in the absence of an AI tool. We have currently attempted to use the AI technologies to create films in order to apply our model to them using datasets of public images. In order to improve the model's performance, we will try to concentrate on this community in our upcoming works, collect a database of data for training, and try to pay attention to the small details that make a difference in the autistic face with medical assistance. We might additionally modify the architecture to make it more consistent and performant to be applied to this kind of population or integrate contextual cues (e.g. body language, vocal tones) to reduce reliance on facial data alone. To achieve even more substantial findings that are aligned with the specificity of these populations and accepted by the medical party, many parties need to work together and concentrate on understanding facial features and emotions.

References

[1] Frith, Uta. (1989). A new look at language and communication in autism. The British journal of disorders of communication. 24. 123-50. 10.3109/13682828909011952.

[2] F. E. E. Rhatassi, B. E. Ghali, and N. Daoudi, “Deep Learning Approaches for Recognizing Facial Emotions on Autistic Patients”

[3] Development of video-based emotion recognition using deep learning with Google Colab published in TELKOMNIKA Vol. 18 No. 5, October 2020.

[4] Jad Haddad, Olivier Lezoray, and Philippe Hamel. 2020. 3D-CNN for Facial Emotion Recognition in Videos. In Advances in Visual Computing: 15th International Symposium, ISVC 2020, San Diego, CA, USA, October 5–7, 2020, Proceedings, Part II. Springer-Verlag, Berlin, Heidelberg, 298–309. https://doi.org/10.1007/978-3-030-64559-5_23

[5] Rani, P.I., Muneeswaran, K. Recognize the facial emotion in video sequences using eye and mouth temporal Gabor features. Multimed Tools Appl 76, 10017–10040 (2017). https://doi.org/10.1007/s11042-016-3592-y

[6] Sunan Li, Wenming Zheng, Yuan Zong, Cheng Lu, Chuangao Tang, Xingxun Jiang, Jiateng Liu, and Wanchuang Xia. 2019. Bi-modality Fusion for Emotion Recognition in the Wild. In 2019 International Conference on Multimodal Interaction (ICMI '19). Association for Computing Machinery, New York, NY, USA, 589–594. https://doi.org/10.1145/3340555.3355719

[7] Muhammad, F.; Hussain, M.; Aboalsamh, H. A Bimodal Emotion Recognition Approach through the Fusion of Electroencephalography and Facial Sequences. Diagnostics 2023, 13, 977. https://doi.org/10.3390/diagnostics13050977

[8] Do, LN., Yang, HJ., Nguyen, HD. et al. Deep neural network-based fusion model for emotion recognition using visual data. J Supercomput 77, 10773–10790 (2021). https://doi.org/10.1007/s11227-021-03690-y

[9] C. Gan, J. Yao, S. Ma, Z. Zhang, and L. Zhu, “The deep spatiotemporal network with dual-flow fusion for video-oriented facial expression recognition,” Digital Communications and Networks, vol. 9, no. 6, pp. 1441–1447, Dec. 2023, doi: 10.1016/j.dcan.2022.07.009.

[10] Hajarolasvadi, Noushin & Demirel, Hasan. (2020). Deep Facial Emotion Recognition in Video Using Eigenframes. Image Processing, IET. 10.1049/iet-ipr.2019.1566.

[11] N. Samadiani, G. Huang, Y. Hu and X. Li, "Happy Emotion Recognition From Unconstrained Videos Using 3D Hybrid Deep Features," in IEEE Access, vol. 9, pp. 35524-35538, 2021, doi: 10.1109/ACCESS.2021.3061744.

[12] Chouhayebi, H.; Mahraz, M.A.; Riffi, J.; Tairi, H.; Alioua, N. Human Emotion Recognition Based on Spatio-Temporal Facial Features Using HOG-HOF and VGG-LSTM. Computers 2024, 13, 101. https://doi.org/10.3390/computers13040101

[13] Chen, Y., Li, J., Shan, S., Wang, M., & Hong, R. (2024). From static to dynamic: Adapting Landmark-Aware image models for facial expression recognition in videos. IEEE Transactions on Affective Computing, 1–15. https://doi.org/10.1109/taffc.2024.3453443

[14] Y. Ouzar, F. Bousefsaf, D. Djeldjli and C. Maaoui, "Video-based multimodal spontaneous emotion recognition using facial expressions and physiological signals," 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), New Orleans, LA, USA, 2022, pp. 2459-2468, doi: 10.1109/CVPRW56347.2022.00275.

[15] Naveen N.C., Sai Smaran K.S., Shamitha A.S (2024). Real Time Facial Emotion Recognition using Deep Learning Models,"International Journal of Computer Applications (0975 – 8887) Volume 186 – No.29

[16] Li, S., Zheng, W., Zong, Y., Lu, C., Tang, C., Jiang, X., Liu, J., & Xia, W. (2019). Bi-modality Fusion for Emotion Recognition in the Wild. 2019 International Conference on Multimodal Interaction.

[17] N. T. Singh, S. Rana, S. Kumari and Ritu, "Facial Emotion Detection Using Haar Cascade and CNN Algorithm," 2023 International Conference on Circuit Power and Computing Technologies (ICCPCT), Kollam, India, 2023, pp. 931-935, doi: 10.1109/ICCPCT58313.2023.10245125.

[18] K. Zhang, Y. Huang, Y. Du and L. Wang, "Facial Expression Recognition Based on Deep Evolutional Spatial-Temporal Networks," in IEEE Transactions on Image Processing, vol. 26, no. 9, pp. 4193-4203, Sept. 2017, doi: 10.1109/TIP.2017.2689999.

[19] Z. Xia, X. Hong, X. Gao, X. Feng and G. Zhao, "Spatiotemporal Recurrent Convolutional Networks for Recognizing Spontaneous Micro-Expressions," in IEEE Transactions on Multimedia, vol. 22, no. 3, pp. 626-640, March 2020, doi: 10.1109/TMM.2019.2931351.

[20] Fan, Y., Lam, J. C. K., Li, V. O. K., & Kot, A. C. (2020). Frame attention networks for facial expression recognition in videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 44*(9), 6096–6108. https://doi.org/10.1109/TPAMI.2020.2974829

[21] Zhang, X., Wang, Y., Zhang, C., & Liu, C. (2021). Hybrid attention cascade network for facial expression recognition. *Sensors, 21*(6), 2003. https://doi.org/10.3390/s21062003

[22] Talaat, F. M., Abou El-Magd, H., & El-Hosseini, M. A. (2023). Real-time facial emotion recognition system among children with autism based on deep learning and IoT. Neural Computing and Applications, 35(29), 21023–21038.

[23] Abou El-Magd, H., El-Hosseini, M. A., & Talaat, F. M. (2024). Real-time facial emotion recognition model based on kernel autoencoder and convolutional neural network for autism children. *Soft Computing, 28*(6), 14873–14891

[24] Lee, J. H., & Wong, K. W. (2020). AEGIS: A real-time multimodal augmented reality computer vision based system to help individuals with autism spectrum disorders recognize emotions. arXiv.