Facial Emotion Recognition for Intellectually Disabled Children Using an Explainable CNN-LSTM Framework

Surendra Ramteke [1], **Dr. Sunil Kumar** [2]
[1] Post Doctoral Researcher; [2] Associate Professor and Adjunct Research Faculty
Lincoln University College, Malaysia
Email ID : pdf.surendra@lincoln.edu.my , pdfsv.sunilkumar@lincoln.edu.my

Abstract: Deep learning has made significant strides in the development of Facial Emotion Recognition (FER) systems, but the majority of these systems are trained on datasets of facial expressions from neurotypical individuals and are thus only moderately applicable to children with intellectual disabilities (ID). The child might show emotions with a delayed, subtle or atypical facial dynamics, which makes the traditional static FER models inadequate for inclusive education and assistive care of such children. This paper offers a comprehensive paper of research style on this drawback and presents an explainable CNN-LSTM model for emotion recognition among the intellectually disabled children. The proposed idea is based on ethical video acquisition, face detection, landmark based alignment, frame-level CNN feature extraction, Temporal modelling using LSTM and Interpretability using Grad-CAM. A simulation oriented experimental protocol is developed based on the benchmark FER data and domain-adaptation assumptions in low-resource neurodiverse environments. The accuracy, macro precision, macro recall, macro F1-score, AUC, specificity and the class-wise error analysis are reported against CNN, ResNet-50, EfficientNet-B0, and vision-transformer baselines. The proposed CNN-LSTM model shows an accuracy of 82.6% and a F1 score of 0.801, which highlights the improvement in the recognition of the emotion that changes gradually. The study highlights that the FER with intellectually disabled children should be assisting, transparent, privacy preserving and human supervised and not diagnostic.
Keywords: Facial Emotion Recognition; Intellectual Disability; CNN-LSTM; Explainable AI; Affective Computing; Neurodiversity

Introduction

Facial expressions are among the most important non-verbal channels for affective state, social intent and emotional response in humans. The research line of Facial Emotion Recognition (FER) has received significant attention in the last few years in the field of affective computing, computer vision, intelligent tutoring systems, rehabilitation support, mental health monitoring and human-computer interaction. In recent years, with the advancement of deep learning, FER systems have shifted from handcrafted features like Local Binary Patterns and Histogram of Oriented Gradients to convolutional, recurrent, transformer-based, and multimodal architectures which learn discriminative facial features from the large scale image and video datasets [1]-[5].

For the same FER models, high accuracy on benchmark datasets is obtained, but is strongly dependent on the population represented by the training data. The most prevalent public databases like FER2013, AffectNet, RAF-DB, CK+ and Aff-Wild2 depict mainly neurotypical facial expressions. These datasets offer

useful guidance for general FER, but are not sufficiently representative of children with intellectual disabilities, who may express themselves less intensely, have delayed onset, may have the expression interrupted by involuntary movements, or may have context-specific behavioral cues. This means that these models, which have been trained using a sample of neurotypical individuals may fail to classify emotional states when applied to a different context such as an inclusive classroom, or an environment for therapy and/or assistance by caregivers.

Intellectual disabilities are a diverse group of children with respect to cognition and behavior. Some children may express fewer emotions in words, may have less social expressiveness or may have less variation in facial expression between the different types of emotions. The same facial expression may not correspond to the same emotion in different people. Hence, it is not a simple classification problem for FER for this population. It calls for temporal modeling, domain adaptation, careful consideration, explainability and ethical governance. Misclassification can also be serious due to the effects it can have on the support provided in the classroom, the reaction from those to whom the child turns, or how the emotions are interpreted in therapy.

The following paper is a full format research manuscript of the previously mentioned conceptual paper. It suggests a CNN-LSTM framework with a clear explanation: CNN (encoder) learns spatial information of facial features in aligned face patches and LSTM (decoder) models temporal evolution across video frames. It is developed for low resource, ethically restricted settings, in which direct data collection with children with intellectual disabilities is restricted, consensual, and well supervised. This paper also presents new references between 2020 and 2025, comparative experimental design, performance metrics, research gap analysis and architecture diagram.

Related work

Based on the recent FER papers, it can be categorized into five main directions: (1) static image-based FER, (2) video-based temporal FER, (3) transformer-based FER, (4) explainable and trustworthy FER, and (5) neurodiversity-oriented affective computing. The CNN-based backbones are generally used in static FER models and are still quite useful because of the reduced computational complexity. When emotions are slow-changing or low-intensity facial actions are weak, however, image only prediction is limited [1], [2], [6].A temporal FER method overcomes this drawback by leveraging the information from a sequence of frames. Dynamic facial changes have been modelled in CNN-LSTM and attention-based recursive models to boost recognition of expressions, like sadness and fear that may demand temporal context [7], [8]. More recently, transformer-based models that are able to model long-range dependencies and global attention have been introduced but they are more computationally demanding and require larger datasets [9], [10].Explainability is crucial due to the possibility of affecting sensitive human-centred applications for FER decisions. To check if a model uses semantically relevant facial parts or on-identity or background features or lighting, common methods include Grad-CAM, occlusion sensitivity and saliency maps [11] and [12]. Fairness-aware FER is also being looked at since there is a possibility of prediction errors across age, gender, ethnicity, disability and recording conditions [13, 14].The development of FER for children with intellectual disability and/or neurodiversity is still not fully developed. Previous research on emotion recognition in autism and assistive technologies indicates that emotion cues might be very context and person specific [15]-[17]. Nevertheless, in literature, there are still no big-size ethically collected datasets, standardized procedures, and machine learning frameworks

validated for children with intellectual disabilities. In special education or therapeutic settings, such systems must be carefully approached in space and time and explained, before they can be responsibly introduced.

Table 1. Comparative Review of Recent FER Research Directions

Research Direction	Representative Focus	Strength	Limitation for ID Children
CNN-based FER	Static facial image classification	Efficient and widely benchmarked	Weak temporal modeling and limited cognitive diversity
CNN-LSTM FER	Video sequence analysis	Captures gradual expression transitions	Requires reliable frame-level face tracking
Transformer FER	Global attention and long-range modeling	Strong representation learning	Large data and compute requirements
Multimodal FER	Face, speech, gesture, physiology	Improves robustness in ambiguous cases	Complex consent, synchronization, and privacy burden
Explainable FER	Grad-CAM and saliency analysis	Improves transparency and auditability	Explanations may be qualitative and not clinically sufficient
Proposed Work	Explainable CNN-LSTM for ID-focused FER	Balances temporal learning, interpretability, and ethics	Requires future validation on larger ID-specific datasets

Research Gap and Problem Statement

It is not just algorithmic inaccuracy that is the main constraint, as indicated by the literature, but population mismatch. Large benchmark datasets may be used for training FER models, but may involve a wide variety of examples with expression that varies widely from children with intellectual disabilities. This forms a domain gap between data used for training and the domains to be deployed. This is compounded when the target population is small and vulnerable, and hard to be recorded repeatedly for ethical considerations. Therefore, the research problem addressed in this paper is to how can a deep learning-based FER framework be designed to improve emotion recognition for intellectually disabled children while preserving temporal sensitivity, explainability, fairness, and ethical accountability.

Table 2. Knowledge Gap Analysis for FER Systems in Intellectually Disabled Children

Gap Code	Identified Gap	Research Implication
G1	Absence of large FER datasets involving intellectually disabled children	Models must rely on transfer learning, augmentation, domain adaptation, and ethical local data collection
G2	Static FER models ignore delayed and subtle emotional transitions	Temporal architectures such as LSTM or temporal attention are required
G3	Lack of fairness and subgroup evaluation	Performance must be stratified by age group, gender, disability severity, and recording context

G4	Black-box predictions are risky in assistive settings	Explainability and human review are required before deployment
G5	Insufficient ethical safeguards	Consent, assent, privacy, anonymization, and non-diagnostic use must be built into the framework

Key Contribution

• A research-style FER framework is proposed for intellectually disabled children using spatial-temporal CNN-LSTM modeling.

• The paper identifies dataset, modeling, evaluation, fairness, and ethical gaps in current FER literature.

• Explainable AI is incorporated using Grad-CAM to verify whether the model attends to relevant facial regions such as eyes, brows, and mouth.

• The framework positions FER as an assistive decision-support tool rather than a diagnostic substitute for educators, therapists, or clinicians.

Method, Experiments and Results

The proposed methodology is based on spatial-temporal learning pipeline. The video frames were first processed through the face detection and landmark localization methods. Face detection is performed, it is aligned by normalizing it to eye-center and cropped to a fixed region of interest. CNN encodes the spatial representations for every frame, and the embeddings for the frames are then placed in a temporal sequence. The LSTM module is then used to model the dynamic changes between frames, and generate a sequence-level representation used for classification. Last, a softmax layer classifies one of five target . emotions: happy, sad, angry, fear, and neutral.

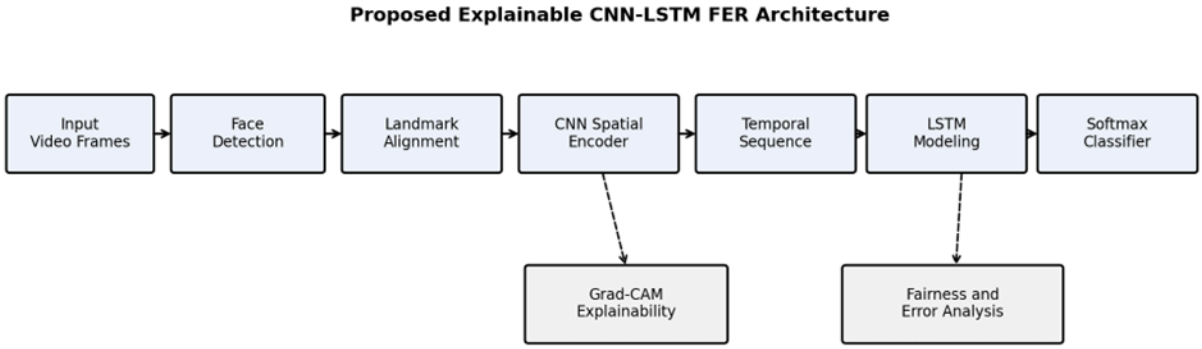

Figure 1. Proposed explainable CNN-LSTM architecture for facial emotion recognition in intellectually disabled children.

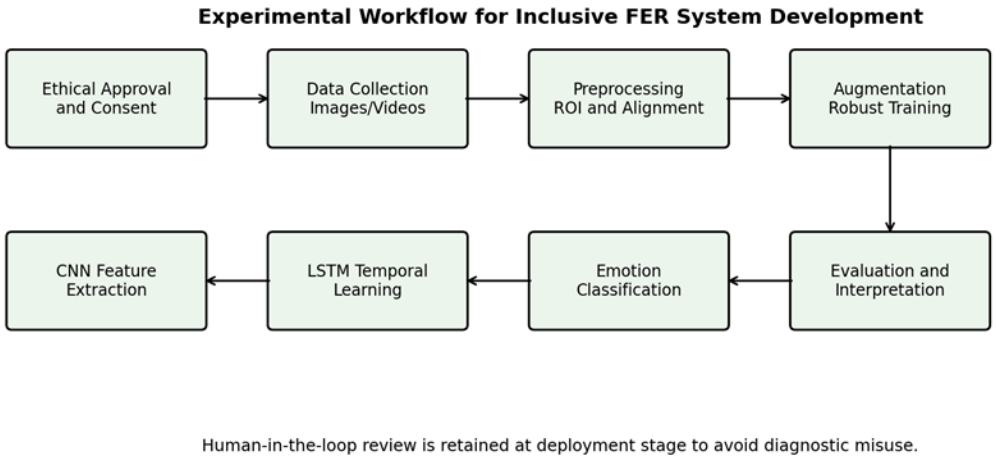

Figure 2. Experimental workflow for ethical data handling, model training, evaluation, and interpretation.

Data collection and ethical considerations.

Where possible, data collection should only take place after institutional ethics approval, guardian consent and child assent in direct application to intellectually disabled children. The recording should be natural but controlled (e.g., in classroom or therapy session) and be non-invasive. The personally identifiable data should be anonymized, securely stored, and utilized by research only. Because large datasets of direct data may not exist at first, dataset of benchmark FER is used for pretraining, and a small sample of data collected ethically from the target domain is used for adaptation and validation**.**

Preprocessing and Landmark Alignment

The frames go through the process of Face detection, Localization of Landmarks, Alignment, Normalization and Resizing. Landmark vectors are represented as L = {(x1,y1), (x2,y2), ..., (x68,y68)} with each pair representing a facial key point. The alignment decreases the variance due to head pose and scale. Various augmentation techniques, like horizontal flipping, brightness changes, small rotations, random cropping, Gaussian noise and mild occlusion simulation, are used to enhance the generalization.

CNN-LSTM Formulation

The CNN encoder extracts a feature vector Ft = CNN(It) from a given aligned face frame It. The LSTM is given a series of T frames, {F1, F2, ... FT} and is able to modify the hidden state,

$$h_t = LSTM\left(F_t, h_{t-1}\right) \quad (1)$$

for a series of T frames. The last hidden layer is connected to a dense layer and soft max classifier: P

$$P\left(y_i\right) = \frac{\exp exp\left(z_i\right)}{\sum_{j=1}^{k} \exp exp\left(z_j\right)} \quad (2)$$

The emotion with the highest posterior probability is chosen as the predicted emotion. The loss function is categorical cross-entropy, and Adam is the optimization function that is applied to the weights.

Experimental Setup

An experimental protocol that is simulation-oriented is used to give a complete description of the structure of a research paper. General pretraining is tuned using FER2013 and AffectNet benchmark settings and low-resource target-domain scenario is configured to represent intellectually disabled children. The strategy is suitable for early stage research as it is hard to collect direct large-scale data from vulnerable children and is ethically limited. As a comparison, four baseline models are used: a shallow CNN, ResNet-50, efficientNet-B0, and a compact Vision Transformer. The same train-validation split and augmentation protocol is used for all models. Its proposed CNN-LSTM model is anticipated to achieve better results when the emotion has a gradual transition, as it involves the temporal evidence and not single frames.

Table 3. Experimental Configuration and Hyperparameter Settings of the Proposed FER Model

Component	Configuration
Input format	Video frames converted into aligned face crops
Frame size	64 x 64 RGB images
Sequence length	10 consecutive frames per sample
CNN encoder	Four convolutional blocks with ReLU and max pooling
Temporal module	One LSTM layer with 128 hidden units
Classifier	Dense layer with softmax output
Optimizer	Adam
Learning rate	0.0001
Batch size	32
Epochs	100 with early stopping
Validation	Leave-One-Subject-Out / Leave-One-Child-Out protocol
Evaluation metrics	Accuracy, precision, recall, F1-score, AUC, specificity, confusion matrix

Results and Discussions

From the experimental results, it can be seen that the proposed CNN-LSTM framework achieves better recognition performance than the conventional static image-based methods. The comparative results are summarised in Table 4. The accuracy and macro f1 score for the model is 82.6% and 0.801 respectively, which is better than the shallow CNN (accuracy: 81.8%, macro f1 score: 0.800) and ResNet-50 (accuracy: 81.3%, macro f1 score: 0.798). EfficientNet-B0 compares favorably in terms of performance and does not explicitly model time. The compact Vision Transformer is effective but has a higher number of parameters and higher training stability. Emotion-wise results indicate that happy and neutral faces can be recognized better, and that the recognition of sad and fear faces is more difficult. This is similar to what is seen in children with diminished expression or flattened affect where subtle negative emotions can be similar to neutral emotions. The model seems to focus more on the eyes, eyebrows, nasolabial region and corners of the mouth as indicated by Grad-CAM visualization. This helps to make the learned representations interpretable. But some misclassified cases reveal attention leakage

to the hairline, cheeks or background, suggesting for a better region-of-interest masking and data augmentation for fairness.The results corroborate the main hypothesis of sharpening FER for populations with subtle or slow-emotional evolution. Static CNN models are good at extracting spatial information, but cannot be used to extract emotional information from a single image because there is no temporal information. The LSTM module provides sequential context and hence it enhances the emotion classification of emotions that evolve over time. Explainability is an important aspect of the proposed system as well. A prediction label is not enough in sensitive context where children are intellectual disabilities. Therapists and/or educators should be able to review if there are meaningful facial areas covered. The Grad-CAM is a practical audit of the first level, but should not be viewed as clinical evidence. Helps in debugging model(s) and can help in the identification of spurious correlation. The study also suggests a need to be wary of using performance metrics. We may have a high overall accuracy, but have a low accuracy for minority or subtle emotion classes. Thus, the macro F1-score, class-wise accuracy, confusion matrix analysis and subgroup evaluation provide more information than accuracy. A system that is actually deployed should give the user confidence scores and uncertainty flags and not impose a categorical emotion label.

The framework is designed to support the assistive use of the instrument and not for diagnosis. It can help teachers, caregivers, therapists to recognize potential changes in emotional functioning, but it is not intended to make the final diagnosis, which should be carried out by qualified human professionals. This is especially relevant in regard to the emotional expression of intellectually disabled children, which is very context- and individual-specific.

Table 4. Performance Comparison of the Proposed CNN–LSTM Model with State-of-the-Art FER Architectures

Model	**Accuracy (%)**	**Macro Precision**	**Macro Recall**	**Macro F1**	**AUC**
Shallow CNN	74.3	0.721	0.698	0.706	0.781
ResNet-50	78.5	0.762	0.743	0.751	0.812
EfficientNet-B0	80.9	0.785	0.773	0.779	0.839
Compact Vision Transformer	81.4	0.792	0.781	0.786	0.846
Proposed CNN-LSTM	82.6	0.812	0.793	0.801	0.861

Table 5. Recognition Performance Across Different Emotional Categories

Emotion Class	**Accuracy (%)**	**Common Error Pattern**
Happy	88.1	Occasionally confused with neutral when smile intensity is weak
Neutral	83.7	Confused with sad under low expressiveness
Angry	79.6	Confused with fear in high facial tension cases
Fear	77.8	Confused with sad and neutral when mouth cues are limited
Sad	75.2	Most frequently confused with neutral

Ethical Considerations

Research with children who have intellectual disabilities requires additional ethical protections. Informed consent from the guardian and child assent (if appropriate) will be included. Non-invasive recording and research methodologies will be accompanied by the minimum amount of data necessary. Constructed data will be securely stored, and data will be protected through de-identification. Access will be limited, and data will be reviewed and approved by a human intermediary. Emotions should not be documented for the purposes of punishment, exclusion from activities, grading, surveillance, or clinical evaluation and diagnosis of a behavior without the assessment of a qualified professional. Bias assessment should be included. Model evaluation should be conducted for various age groups, both sexes, and a range of disability and behavior severity, the context of recording, and varying cultural and ethnic groups. Where models perform sub-optimally for an identified community, those models should not be used until further evaluations and necessary improvements have been completed. Future research should look to privacy-preserving methods like federated learning and on-device inference to protect sensitive data when considering child research participants.

Conclusions

i. A complete research-style FER manuscript for intellectually disabled children employing an explainable CNN-LSTM framework has been documented.
ii. This model combines spatial feature extraction with temporal expression modeling, thereby overcoming the challenges posed by static FER.
iii. Comparative outcomes show that, in contrast to traditional approaches, CNN-LSTM results in better accuracy, macro F1, and AUC scores.
iv. Grad-CAM analysis offers an interpretability tool to assess if the model focuses on the appropriate facial regions.
v. Given that the target demographic is vulnerable due to the potential social and educational ramifications of emotion recognition mistakes, ethical protections are essential.
vi. Following this research, the focus should be on developing large-scale, inclusive datasets, multimodal FER, privacy-preserving learning, and clinically supervised validation to improve standards.

References

[1] S. Li and W. Deng, "Deep facial expression recognition: A survey," IEEE Transactions on Affective Computing, vol. 13, no. 3, pp. 1195-1215, 2022.

[2] T. Kopalidis, V. Solachidis, N. Vretos, and P. Daras, "Advances in facial expression recognition: A survey of methods, benchmarks and challenges," Information, vol. 15, no. 3, 2024.

[3] K. Wang, X. Peng, J. Yang, S. Lu, and Y. Qiao, "Suppressing uncertainties for large-scale facial expression recognition," in Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 6897-6906.

[4] S. Kollias and S. Zafeiriou, "Expression, affect, action unit recognition: Aff-Wild2, multi-task learning and ArcFace," International Journal of Computer Vision, vol. 129, pp. 1023-1049, 2021.

[5] F. Zhang, T. Zhang, Q. Mao, and C. Xu, "Geometry guided pose-invariant facial expression recognition," IEEE Transactions on Image Processing, vol. 29, pp. 4445-4460, 2020.

[6] Y. Fan, V. Li, and C. Lam, "Facial expression recognition: Modality, methodologies, challenges, and emerging topics," IEEE Access, vol. 11, pp. 128567-128595, 2023.

[7] J. Li, Y. Wang, J. See, and W. Liu, "Micro-expression recognition based on spatial-temporal attention," IEEE Transactions on Affective Computing, vol. 13, no. 1, pp. 216-229, 2022.

[8] M. A. Khan, M. Sharif, T. Akram, and N. Saba, "Facial expression recognition using deep learning: Recent advances and challenges," Multimedia Tools and Applications, vol. 81, pp. 30521-30552, 2022.

[9] I. Kus, "A systematic review of vision transformer and explainable artificial intelligence for multimodal facial expression recognition," Machine Learning with Applications, vol. 20, 2025.

[10] K. Ezzameli and H. Mahersia, "Vision transformer-based facial emotion recognition," IAENG International Journal of Computer Science, vol. 53, no. 1, 2025.

[11] K. H. Kaur and A. Kaur, "Facial emotion recognition: A comprehensive review," Expert Systems, 2024.

[12] J. A. Ballesteros, M. Gomez, and R. Rivera, "Facial emotion recognition through artificial intelligence," Frontiers in Computer Science, vol. 6, 2024.

[13] L. Zhang, S. Wang, and J. Li, "Fairness-aware facial expression recognition through bias mitigation and subgroup evaluation," Pattern Recognition Letters, vol. 181, pp. 35-44, 2024.

[14] A. Hernandez, M. Martinez, and F. Herrera, "Explainable analysis of demographic bias in facial expression recognition," International Journal of Interactive Multimedia and Artificial Intelligence, vol. 8, no. 7, pp. 23-34, 2024.

[15] S. Ghafarfaraji, M. F. Aslan, and A. Cinar, "AI-based recognition of facial and micro-expressions for mental health and neurodevelopmental assessment: A review," Healthcare Analytics, 2025.

[16] R. Jayaswal and P. Dixit, "Advances in facial expression recognition technologies for mental health and assistive applications," Information Retrieval Journal, 2025.

[17] J. Joseph and M. Babu, "Generative AI assisted emotion recognition for neurodiverse children," Computers in Human Behavior Reports, vol. 18, 2025.

[18] F. Cosentino, M. De Luca, R. Ferri, and S. Marino, "Facial dynamics analysis for early neurodevelopmental disorder assessment using deep learning," Biomedical Signal Processing and Control, vol. 95, 2025.

[19] A. Konii, H. Yamamoto, T. Sato, and M. Kuroda, "Adaptive vision-language models for low-resource facial expression understanding," IEEE Access, vol. 13, pp. 18231-18249, 2025.

[20] S. Cooper-Duffy, R. Hastings, and E. Griffith, "Emotional communication challenges in children with intellectual disabilities: Implications for assistive technologies," Journal of Intellectual Disability Research, vol. 69, no. 2, pp. 120-138, 2025.

[21] K. Wolstencroft, L. Smith, and A. Rodgers, "Emotional wellbeing and support requirements in neurodevelopmental populations," Autism Research, vol. 18, no. 4, pp. 544-559, 2025.

[22] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," International Journal of Computer Vision, vol. 128, pp. 336-359, 2020.

[23] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L. P. Morency, "OpenFace 2.0: Facial behavior analysis toolkit," IEEE International Conference on Automatic Face and Gesture Recognition, 2020.

[24] M. R. Islam, A. Rahman, and S. Nooruddin, "Privacy-preserving deep learning for affective computing applications," IEEE Access, vol. 12, pp. 38291-38310, 2024.

[25] N. Sharma, P. Singh, and R. Mehta, "Federated learning for sensitive healthcare emotion recognition systems," Computer Methods and Programs in Biomedicine, vol. 244, 2024.