

RESEARCH ARTICLE

Fused YOLO and Traditional Features for Emotion Recognition From Facial Images of Tamil and Russian Speaking Children: A Cross-Cultural Study

A. MARY MEKALA¹, M. VARALAKSHMI¹, C. P. ACHYUTHA GOWDA²,
LETI MANISH KUMAR², ELENA E. LYAKSO³, OLGA FROLOVA³,
AND RUBAN NERISSON⁴

¹School of Computer Science and Engineering, Vellore Institute of Technology, Vellore 632014, India

²BMS Institute of Technology and Management, Bengaluru, Karnataka 560064, India

³Department of Higher Nervous Activity and Psychophysiology, Saint Petersburg State University, 199034 Saint Petersburg, Russia

⁴School of Electrical Engineering, Vellore Institute of Technology, Vellore 632014, India

Corresponding author: Ruban Nersisson (nruban@vit.ac.in)

This work was supported in part by the Department of Science and Technology, Government of India, under Grant DST/INT/RUS/RSF/P-57/2021; and in part by Russian Science Foundation under Project 22-45-02007.

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Institutional Ethical Committee for Studies on Human Subjects (VIT Vellore) under Application No. VIT/IECH/XIII/2023/16.

ABSTRACT Cross-cultural study that avoids bias in the formulation of emotion recognition models is indispensable to address the challenges in facial emotion classification for children, for it being a relatively difficult task. With evidences from literature for the hybrid feature extraction approaches to improve the image classification accuracy, this research work focuses on developing a hybrid framework for emotion recognition from facial images of Tamil and Russian children. The dataset is audio video recording of 28 Tamil speaking and 64 Russian speaking children. The data is collected in a controlled environment and labelled by experts. Traditional features like Grey Level Cooccurrence Matrix (GLCM) and facial landmark are extracted and are fused with You Look Only Once (YOLO V5) features. While facial landmarks and GLCM provide useful information about the facial expressions and texture of the image, YOLO V5 being a single-stage object detector makes the hybrid model super-fast and achieve high accuracy in detecting small objects and in low-light settings. The different classifiers that includes KNN, SVM, Random Forest, XGBoost, and Multilayer Perceptron, is employed yielding the accuracy result of 93%, 86%, 89%, 88%, and 90%. The use of majority voting ensemble of heterogeneous classifiers for the final prediction strengthens the model further, yielding an accuracy as high as 96% for the custom cross-cultural dataset, consisting of facial images of Russian and Indian children. This is consistent with the results obtained for Indian and Russian datasets. Further, the ablation study unveils the effect of feature fusion in boosting the performance and the dominance of YOLO V5 features over the other two.

INDEX TERMS Facial emotion recognition, YOLO V5, gray level co-occurrence matrix (GLCM), facial landmarks, feature fusion, ensemble classification.

I. INTRODUCTION

Emotions are human reactions made in response to an event or situation that they experience. Emotion Recognition is a

The associate editor coordinating the review of this manuscript and approving it for publication was Sotirios Goudos¹.

field of study that focuses on identifying and interpreting human emotions. This involves the usage of various data sources such as text, speech, facial expressions, or physiological signals [1]. Increased human-computer interaction and widespread use of behavioural biometric systems has drawn the attention of researchers towards the development

of automatic facial emotion recognition systems. Emotional analytics finds its application in various domains including but not limited to road and traffic safety, marketing and healthcare. In the field of road safety, the tool can help to avoid road accidents by detecting fatigue in drivers and raising an alert in real time. In a similar way, AI-powered facial emotion recognition systems are used by market research companies to automate the analysis of video feeds of the user interacting with a product and thereby predict customers' future buying plans. In the healthcare sector, face emotion recognition capability incorporated into care taking robots can be used for patient monitoring. Furthermore, it can help in the early detection of autism in children, as these children have peculiar patterns in their facial expressions.

The existing literature catalogues a broad spectrum of Machine Learning (ML) and Deep Learning (DL) algorithms for emotion classification from facial images; however, even the most sophisticated techniques are not very efficient for predicting the emotions from the facial images of children and for classifying micro expressions [2]. Emotion recognition for children is a relatively difficult task as the children are less expressive for certain emotions and demonstrate only a limited set of emotions while interacting with strangers than with their peers. This necessitates the design and development of new algorithms for emotion classification. Therefore, this research work focuses on developing a hybrid YOLO algorithm for emotion recognition from facial images of Tamil and Russian children. In all the emotion detection approaches studied so far, the different algorithms are tested with openly available benchmark datasets.

The novelty of this research work is that a custom cross-cultural dataset is developed with the typically developing Russian and Indian children that includes only the four major emotions that are of particular interest to our research. Performing cross-cultural research [3], [4], and [5] in which samples are drawn from across the globe to analyse the universal and locale-specific patterns in facial emotion recognition for children, will help to avoid bias in the formulation of emotion recognition models [6], [7].

The study reveals promising results for care takers, therapists, and for the educators to observe the emotional development of children. For the early detection of emotional disorders and delayed development of emotional manifestation in typically developing children. Misunderstanding of children's emotion from diverse cultures may lead to unsuitable responses which can be overcome by the developed approach. The major social crisis and very sensitive issue of today's modern world is, understanding the children effectively and recognizing the manifestation of their emotions, the developed algorithm will contribute to this problem in positive way.

The rest of the paper is organized as follows. Section II reviews some of the related works on hybrid models that employ feature fusion for enhanced image classification. Section III explains the proposed model that integrates the GLCM features and facial landmark features with the

features extracted from YOLO V5 and elaborates the data collection process carried out to build the cross-cultural dataset. Section IV discusses in detail, the performance of the proposed model on three different datasets. This section also compares the results of various classifiers applied to the fused data. Section V concludes research findings and provides guidelines for future extensions of the work.

II. LITERATURE SURVEY

Extraction of face from a given image is crucial for the successful implementation of emotion recognition models. This necessitates the selection of the best object detection methods, as the first step. In a study related to object detection, four different methods, like multitask cascaded convolutional networks (MTCNN), Viola-Jones, single-shot multi-box detectors (SSD), and YOLO, have been compared in terms of accuracy and speed. Comparatively, the YOLO algorithm is found to be the best-suited face detection algorithm for its accuracy and speed [8]. YOLO V7 is capable of recognizing the obscured faces and provides a better accuracy in real time [9]. YOLO v5 with attention mechanism has also been successfully applied for the detection of pedestrians in unsupervised railway tracks. Four salient features - Distance Intersection over Union (DIoU) loss for the better prediction of bounding boxes, context and content attention modules and L1 regularization along with batch normalization are implemented in this study to improve the detection accuracy [10]. Detection and localization of the key points on the face is the next important step for any facial image processing task including biometric recognition and facial expression recognition. More advanced techniques for facial landmarks localization have been developed in the recent times [11]. An additional contribution to this domain is the use of the latest facial landmark detection algorithms in addition to the conventional edge and corner detection algorithms to locate eyes, eyebrows, lips and nose and used the generated feature vectors as input to the MultiLayer Perceptron (MLP) for classification of facial expressions [12]. Calculating the gray level co-occurrence matrix (GLCM) has also been found to be useful in facial retrieval where the GLCM value of an unknown gray face image is compared with those of the images stored in the database system resulting in a high degree of matching accuracy [13].

Several machine learning (ML) and deep learning (DL) methods are reported in the literature for the task of facial emotion classification. While ML models require more preprocessing time, Deep learning models consume more time for training and testing.

One such ML model is the Support Vector Machine (SVM) framework proposed for classifying facial emotions from the features extracted using the region-based ORB and Local Binary Patterns (LBP) algorithms. Among the three databases- the Cohn-Kanade database (CK+), MMI database, and Japanese Female Facial Expressions

database (JAFFE) used for verifying the efficacy of the method, highest accuracy is observed for the MMI database [14]. Another comparative analysis between SVM and kNN for the identification of facial expressions using the features extracted from Local Monotonic Pattern (LMP) and Gray Level Co-occurrence Matrix (GLCM) has recorded an accuracy of 93% for SVM [15]. An experimental study to analyse the various feature extraction methods such as Gabor filters, Local Binary Pattern (LBP) and Histogram of Oriented Gradients (HOG) and machine learning algorithms including k Nearest Neighbours (kNN), Random Forest and Support Vector Machine for emotion intensity recognition from both spontaneous and posed facial expressions shows that SVM with LBP feature extraction outperforms the other models [16].

Facial action coding system that defines each micro change in the face muscle as an action unit and assigns a distinct number to each, helps to determine those facial muscles involved in a specific facial expression. These action units are capable of detecting all the basic facial emotions. A comparative study of several ML models including Random Forest, Decision tree and alike revealed that SVM and Logistic regression performed well in the emotion prediction using the features extracted using Histogram of Oriented Gradients (HOG) [17]. Another comparative study evaluates the emotion detection accuracy of MLP and 5 different machine learning algorithms including Naive Bayes (NB), Random Forest, kNN, SVM and Logistic Regression for 3 datasets - Japanese Female Facial Expression (JAFFE), Cohn-Kanade (CK+) and Real-world Affective Faces Database (RAF-DB). They employ Super Resolution GAN for preprocessing and MediaPipe for generating the key points on the face images. A higher accuracy is recorded for Happy, Angry and Surprise emotions compared to the other emotions [18].

With the advent of deep neural networks, a Convolutional Neural Network (CNN) has been used to recognize the facial emotions, recording an accuracy of 95% on FER-2013 dataset and 79% on extended Cohn-Kanade (CK+) datasets [19]. A customized VGG architecture is proposed to overcome the vanishing gradient effect and to improve the overall performance of the facial emotion recognition system [20]. Deep neural networks such as MTCNN and Xception models are also employed for improving the quality of the input images and thereby improving the performance of Face Emotion Recognition (FER) models [21]. Hybrid architectures involving CNN and other neural networks have been found to be more efficient in face emotion recognition. A CNN-LSTM-based neural network has been proposed for identifying facial emotions. Accuracy scores of 78.52% and 63.35% are obtained for the CREMAD and RAVDEES datasets respectively [22]. A similar study on hybrid approach with CNN and RNN to detect the facial emotions using the temporal and contextual intricacies obtained from video recordings of the Emotional Wearable Dataset 2020 results in an accuracy of 63% for test data [23].

Another hybrid variant with CNN-RNN for face emotion recognition uses the features extracted using three methods: Texton, Bag of Words (BOW), and GLCM. The proposed architecture gives an accuracy of 95% while finding the facial emotions like angry, disgust, fear, happiness, sadness, and surprise [24]. Following a similar trend, a hybrid CNN with a deep belief network (DBN) is introduced to identify the facial emotions. The DBN combines the segment-level spatial and temporal features to improve its performance over the existing methods [25].

In spite of the advancements in image classification techniques, images captured in low lighting conditions pose a challenge to facial emotion recognition. In order to combat this challenge, thermal images are also experimented. Use of thermal images, focus given only to specific regions of the face and parallelization of the training and testing phases have all helped to increase the speed and classification accuracy of CNN [26]. Furthermore, in an effort to build a more robust thermal image processing algorithm, an approach that works on a sequence of face images instead of single images has been proposed. GLCM features are extracted, and a parallel deep emotion net is used to make a more reliable classification of a given sequence of images [27].

The first hybrid approach relevant to the proposed work is the integration of Gabor features with the features extracted by CNN, as CNN may fail to focus on minute changes in the facial expressions. It employs an attention module to extract the most important features that helps to improve emotion recognition accuracy [28]. In another feature fusion experiment the CNN is used for classifying seven basic facial expressions: angry, disgust, fear, happiness, neutrality, sadness, and surprise. It focuses on fusing the features extracted using region-based oriented FAST and rotated BRIEF (ORB), and Local Binary Pattern (LBP), and convolutional neural network (CNN) from facial expression images, which has resulted in a very high accuracy score of 98.13% [29]. In an attempt to perform early detection of skin cancer, the features extracted by You Only Look Once (YOLO v2) are fused with the texture and colour features extracted using GLCM and Gabor methods. The fusion network results in an improved classification score of 94% [30]. Another feature fusion-based research focuses on using facial landmarks along with Hu's moments, GLCM and color histograms to train 4 machine learning models namely Random Forest, Linear Regression, KNN and artificial neural networks to assess the facial beauty among which KNN performs the best [31].

Therefore, this research work focuses on developing a hybrid YOLO algorithm for emotion recognition from facial images of Tamil and Russian children. In all the emotion detection approaches studied so far, the different algorithms are tested with popular datasets available openly. The novelty of this research work is that a cross-cultural dataset is developed with the typically developing Russian and

Indian children. This offers various options for exploring different algorithms with different datasets.

III. METHODOLOGY

From various facial emotions, the four emotions that are considered for this research study are: Happy, Neutral, Angry, and Sad.

A. DATA COLLECTION

The Child Emotional Development Method (CEDM) [32] is designed to assess the emotional development of children by determining their ability to express their own emotions. The methodology depends on the children's age, language, and culture of the country. The happy emotion is mostly similar across cultures and can be identified universally, But the emotions like anger, disgust and fear may differ across the different cultures. It is mainly due to the cultural factors that are their facial expressions they show when they interact with people including their body language [33], [34], [35]. The video recordings of Russian children's emotional expressions were generated in the laboratory. Sony HDR-CX560 video camera (maximum resolution 1920×1080 at 50 frames per second) is used to record the facial expression. The Indian children's video recording is generated using a Nikon D3500 digital camera (maximum resolution: 1920×1080 at 50 frames per second). The parents of the children involved in the study, signed an informed consent form approved by the Ethical Committee of St. Petersburg State University for Russian Children and Vellore Institute of Technology for the Indian Tamil-speaking children.

B. DATASET

The participants for this study comprise 28 Indian Tamil-speaking children (10 boys and 18 girls) and 64 Russian children (32 boys and 32 girls). The considered age group for both Indian and Russian children is between 5 and 16 years. The camera is placed at a 1 meter gap from the child's face. The children who participated in the study portrayed the emotions of happy, neutral (calm), sad, and angry through facial expressions. Then the long video is fragmented in coordination with the experts. Individually, for Indian Children, 80 fragments (40 boys and 40 girls) and for the Russian Children, 45 fragments (27 boys and 27 girls) were created. For identification purposes, before each video fragment, a unique number is assigned. The time duration of individual fragments is between 3 and 4 seconds, with a 5-second pause between fragments. The overall duration of each test is approximately 2-3 minutes. The Pinnacle Studio 1.0.0.155 video editing tool is used to fragment the videos. Totally 1141 (Tamil girls - 349 frames; Tamil boys - 269 frames; Russian boys - 259 frames; Russian girls - 264 frames) frames were generated from the video.

A perceptual study was conducted using 10 Indian and Russian experts each. The videos of the Indian and Russian children were given to both the experts to manually recognize the children's emotional state: Happy, Neutral, Sad,

Angry [36]. The results of the perceptual study are cross validated by experts and pediatricians at both sides and the labelling on the data is finalized based on the validation.

The individual frames are normalized and then denoised using OpenCV library function. It implements non-local means denoising algorithm that considers a small window around the noisy pixel and replaces the pixel with the average of all the other windows similar to this window. This denoising technique yields better results than the other blurring techniques.

C. EXPERIMENTAL SETUP

The hardware configuration includes CPU: Intel(R) Core(TM) i9-10900F (2.80 GHz), GPU: NVIDIA GeForce RTX 3090, VRAM: 24 GB, Shared memory: 15.95 GB, Operating System: 64-bit operating system, x64-based processor, and the software configuration setup are Programming Language: Python 3.12.9, CUDA Version: CUDA 12.3, IDE: Jupyter notebook 7.3.2.

D. PROPOSED MODEL

You Only Look Once (YOLO) is super-fast as it processes the entire image in a single pass and hence can be used in real time. In particular, YOLO V5 achieves high accuracy in detecting small objects and in low-light settings. Thus, YOLO V5 is used for the construction of the proposed hybrid model. The performance metrics used to assess YOLO depends on the local system configuration. The parameter average inference time achieved on the local system is 42.51ms. Fig. 1 shows the schematic of the proposed architecture that fuses three different sets of features - facial landmark features, Grey Level Co-occurrence Matrix (GLCM) features and the features extracted by the YOLO V5.

The model comprises the following five stages.

- 1) Frames Extraction
- 2) Facial Landmark Localization
- 3) GLCM Feature Computation
- 4) YOLO v5 Feature Extraction
- 5) Feature Fusion and Heterogeneous Ensemble Classification

Each stage contributes to capturing different aspects of facial expressions, unveiling subtle insights from images.

1) FRAMES EXTRACTION

OpenCV is a python library that can be used to access and manipulate video streams and images. The VideoCapture class of OpenCV serves to open a video file or capture videos through a webcam after which any desired operation can be performed on the video. Subsequently, the read() function of the class is used to read and extract the individual frames from the video. Frame rate expressed in terms of "frames-per-second" (fps) determines the number of frames used up in one second of video. The frame selection strategy determines which frames are included. Frames can be chosen at regular intervals, such as every N^{th} frame,

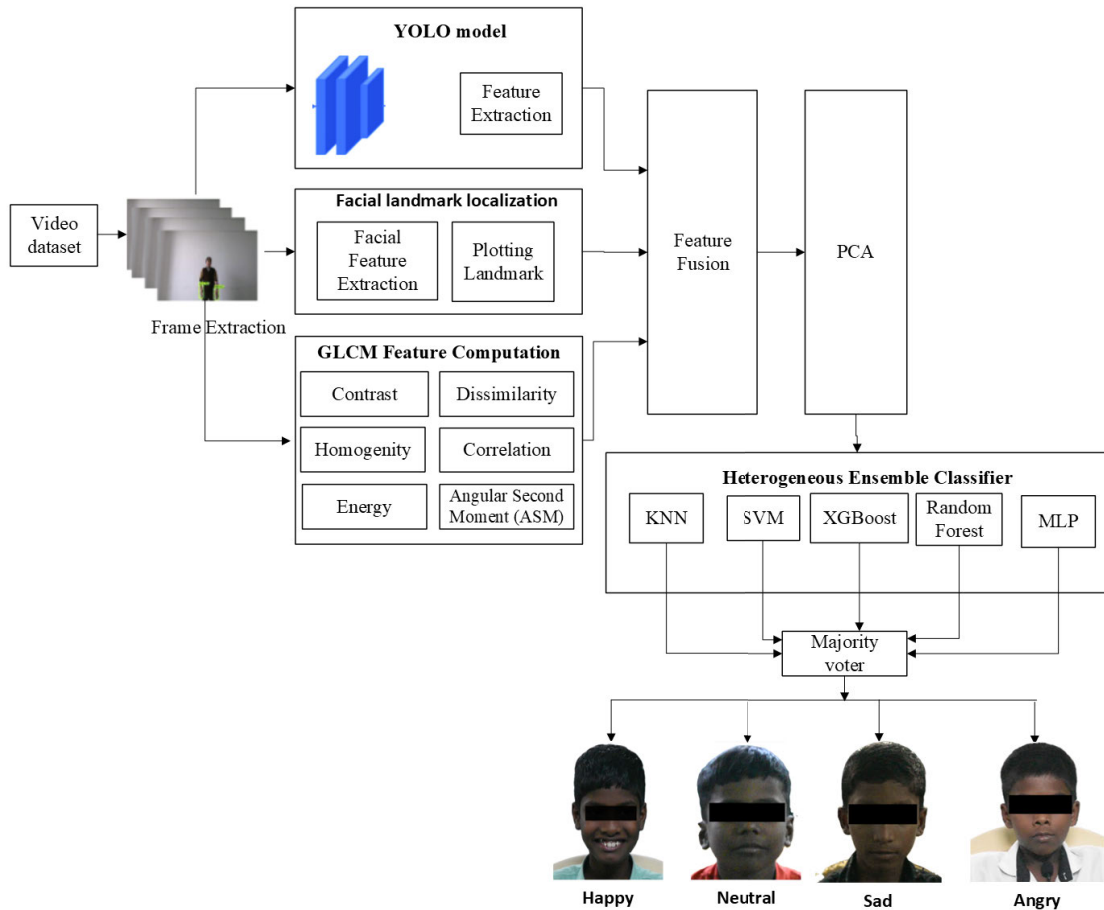


FIGURE 1. Schematic of the proposed architecture.

maintaining a balance between retaining essential information and managing computational load. For instance, consider a video with a frame rate of 30 frames per second. With the frame selection strategy that chooses every 10th frame, a total of 3 frames are selected for every one-second video a single frame. These extracted frames are analysed manually and the frames with low noise and clearly visible facial emotions are selected for further analysis. Four basic emotions namely Happy, Angry, Sad, and Neutral are considered for this study on emotion recognition and so the individual frames are labelled with the appropriate emotion among the four.

2) FACIAL LANDMARK LOCALIZATION

It involves detecting and localizing key points or landmarks on a face, such as nose, eyes, mouth. These landmarks provide useful information about facial expressions and gestures and so commonly employed in emotion recognition in addition to face alignment, facial feature extraction, facial animation and virtual avatar creation. The proposed model uses the Mediapipe library from Google to detect the facial landmarks. It uses Face mesh and other machine learning models to generate a total of 468 three-dimensional landmark points.

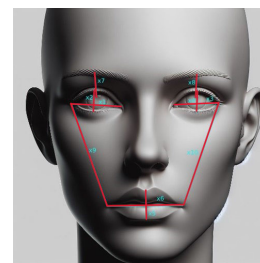


FIGURE 2. Landmark features plotted on a mannequin face image.

Out of these landmarks, 20 specific landmarks are considered to extract the 10 most important features listed in Table 1. Fig. 2 shows the drawing of these landmarks on a facial image.

3) GRAY-LEVEL CO-OCCURRENCE MATRIX (GLCM) FEATURE EXTRACTION

It is a statistical method for image texture analysis. It outputs a matrix that stores the count of pixel pairs with specific values and specific spatial relationship occurring in an image. Several statistical features can be extracted from the GLCM

TABLE 1. Land mark features extracted.

Sl.No.	Description	Symbol
1	Right eye Width	X1=d(173,161)
2	Right eye Height	X2=d(145,28)
3	Left eye width	X3=d(398,388)
4	Left eye height	X4=d(374,258)
5	Lip width	X5=d(62,308)
6	Distance between Lips	X6=d(17,12)
7	Right eye to eyebrow	X7=d(159,52)
8	Left eye to eye brow	X8=d(386,282)
9	Right Lip to eye	X9=d(43,113)
10	Left Lip to eye	X10=d(273,446)

TABLE 2. Calculation of statistical features from grey-level co-occurrence matrix.

Feature	Description	Formulae
Contrast	It measures the spatial frequency of an image and is a different moment of GLCM.	$\sum_i \sum_j i - j ^2 p(i, j)$
Homogeneity	Returns a value that measures the closeness of the distribution of elements in the GLCM to the GLCM diagonal.	$\sum_i \sum_j \frac{1}{1+ i-j ^2} p(i, j)$
Energy	It Returns the sum of squared elements in the GLCM.	$\sum_i \sum_j p(i, j)^2$
Dissimilarity	It is a measure of distance between pairs of objects (pixels) in the region of interest.	$\sum_i \sum_j i - j p(i, j)$
Correlation	Measures the joint probability occurrence of the specified pixel pairs.	$\sum_i \sum_j \frac{(i-\mu_i)(j-\mu_j)p(i, j)}{\sigma_i \sigma_j}$
Angular Second Moment (ASM)	Measures the number of repeated pairs.	$\sum_{i=1}^{N_g} \sum_{j=1}^{N_g} p(i, j)^2$

matrix such as Contrast, Homogeneity, Angular Second Moment (ASM), Energy, Correlation and Dissimilarity. Table 2 enumerates the formulae to calculate the different features. These are calculated for four directions namely horizontal (0°), vertical (90°) and diagonals (45° and 135°) to provide valuable information about the texture of an image.

Here, P(i, j) represents the normalized GLCM. i and j typically represent the intensity values of two neighbouring pixels in the image. μ_i and μ_j are the means of the distributions of p_i and p_j . σ_i and σ_j are the standard deviations of the distributions of p_i and p_j . N is the number of grey levels in the image.

4) YOLO V5 FEATURE EXTRACTION

YOLO V5 typically employs a deep neural network CSP-Darknet53 as its backbone for feature extraction as shown in Fig.3. The trained YOLO V5 model learns the patterns of the facial expressions in the labelled data and gives the coordinates of the bounding box and confidence score as the output. Fig. 4 depicts the process of extracting features from the YOLO V5 model.

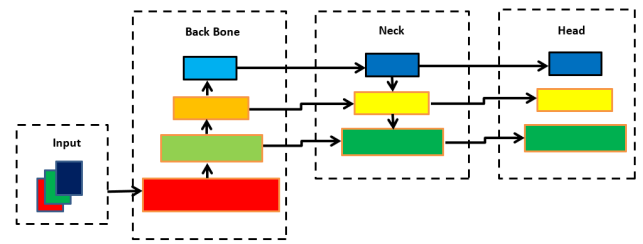


FIGURE 3. YOLO V5 architecture.

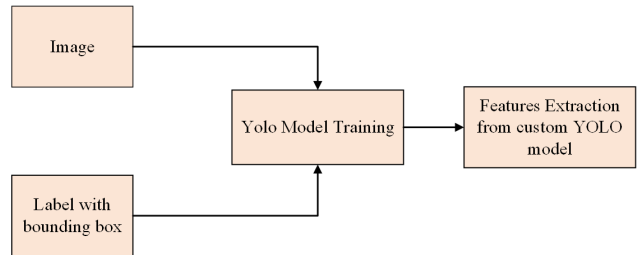


FIGURE 4. YOLO V5 feature extraction.

TABLE 3. Integer representation of emotion class for YOLO labelling.

Value	0	1	2	3
Class	Happy	Sad	Neutral	Angry

Data Annotation and YOLO V5 Training: Label Studio is an open-source data labelling platform used to prepare labelled data for training. This interface is used in our study to quickly and easily draw bounding boxes for the input images and annotate them with emotion classes. Annotations generated by the tool are exported to the YOLO image annotation format that contains the following three details.

- 1) Object-class, an integer representing the emotion class of the image
- 2) Bounding box center coordinates (X center, Y center) normalized by image width and height
- 3) Width and height of the bounding box normalized by image width and height

Fig. 5 illustrates the generation of the emotion class labels and bounding box dimensions for the given images in Label Studio. The integer values assigned to each emotion class for input labelling and the annotations prepared in the YOLO format for a few sample images are presented in Table 3 and Table 4 respectively.

The YOLO formatted input prepared using Label Studio is fed into the YOLO V5 model for training, to predict the emotion class. Features are extracted from the penultimate layer of the trained model, for each image.

5) FEATURE FUSION AND HETEROGENEOUS ENSEMBLE CLASSIFICATION

The hybrid model is proposed to address the difficulty in recognising the emotions of children. The three sets of features including facial landmarks, GLCM and YOLO V5

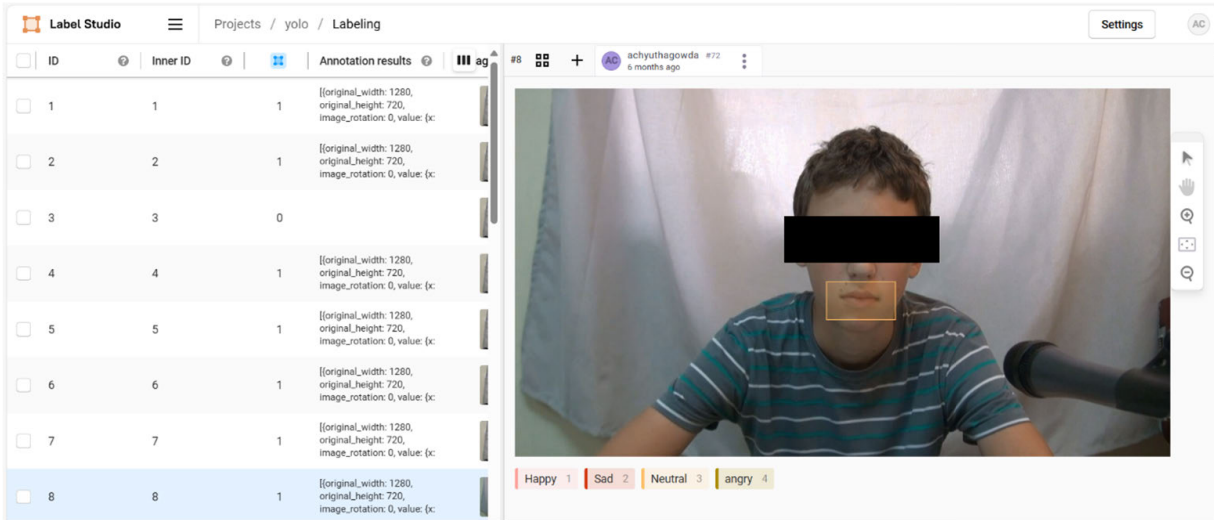


FIGURE 5. Label-studio annotation tool.

TABLE 4. Annotations for sample images.

Image No.	Class	X center	Y center	Width	Height
1	1	0.5183727	0.56926218	0.09711286	0.09798775
2	3	0.50459318	0.88072324	0.16666667	0.16097988
3	1	0.52690289	0.57392826	0.10629921	0.10731992
4	0	0.45275591	0.69291339	0.11286089	0.111986
5	3	0.53674541	0.53193351	0.11811024	0.13531642

features are concatenated. For the next stage of multinomial classification, these concatenated/fused features are fed as input to a heterogeneous ensemble of five different classification models including K-Nearest Neighbours (KNN), Support Vector Machine (SVM), Random Forest, XGBoost and a fully connected feed forward neural network. The parameters considered for hyperparameter tuning for each model are as follows: KNN - n_neighbors=1; SVM - c=1 (regularization parameter); RF - n_estimators = 100; XGboost - random state =42; MLP classifier -learning rate = 0.001, max_iter is adjusted according to the convergence.

As a standard practice, majority voting strategy is applied, in which the class predicted by the maximum number of models is reported as the final result. With feature fusion, the proposed architecture renders itself as a more robust model by being able to handle various types of noise and uncertainties in the data and thus able to generalize well to new and unforeseen input conditions. In addition to the feature fusion, the effective ensembling of different classifiers for the multinomial classification at the final stage, improves the classification accuracy further.

IV. RESULTS AND DISCUSSIONS

A. RESULTS:CLASSIFICATION WITH FUSED FEATURES

For experimental purposes, the dataset consisting of video samples of both boys and girls of Russian and Indian origin, is divided into 3 subsets. The first subset consists of only

TABLE 5. Ten important facial land mark features extracted for a sample image.

X1	X2	X3	X4	X5
45.39824	22.13594	45.04442	22.80351	107.0747
X6	X7	X8	X9	X10
51.0392	24.59675	25.55386	139.2157	130.2997

TABLE 6. GLCM features extracted for a sample image.

Features	0°	45°	90°	135°
Contrast	8.233683	7.728426	14.82665	10.44636
Dissimilarity	1.567757	1.38902	1.969749	1.802938
Homogeneity	0.548836	0.597228	0.502989	0.504364
Energy	0.085067	0.093949	0.08005	0.079577
Correlation	0.9932	0.993617	0.987752	0.99137
ASM	0.007236	0.008826	0.006408	0.006332

Indian subjects, similarly the second subset consists of only Russian subjects and the third subset consists of both Indian and Russian subjects. These various datasets facilitate a comprehensive analysis of displayed emotions. In our study on facial emotion detection, facial landmarks are leveraged to extract the essential information about facial expressions and gestures. Specifically, ten important facial features are extracted from these landmark points, which play a crucial role in our facial emotion detection system. Table 5 shows the 10 facial features extracted for a sample image. Likewise, the features are extracted for each of the individual images in the dataset.

GLCM features provide useful information about the texture of an image. Table 6 presents the important statistical features extracted from the GLCM matrix, computed at four different angles for a sample image.

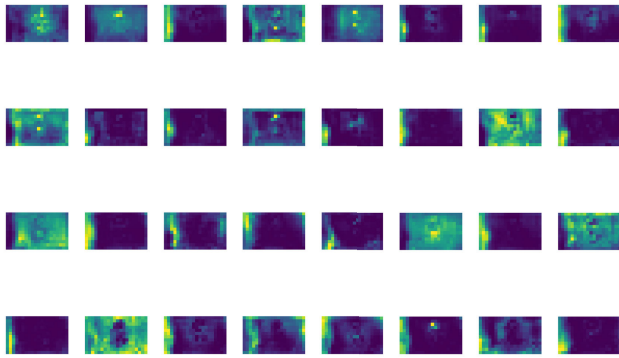


FIGURE 6. Features extracted from the penultimate layer of YOLO V5.

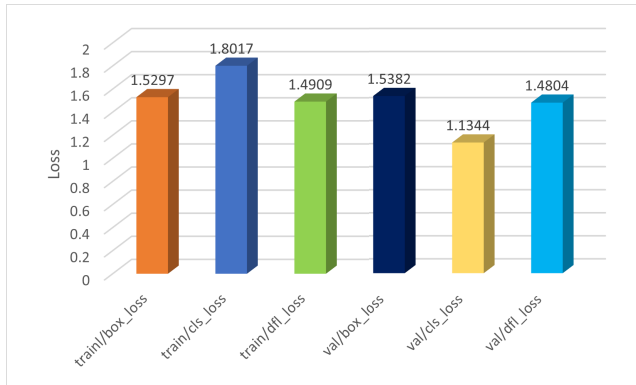


FIGURE 7. Graph showing the Final Values of YOLO Training Parameters.

Features extracted from the penultimate layer of the custom trained YOLO V5 model are an integral part of the proposed hybrid model input and Fig. 6 plots those features extracted for a sample image. The training scores of the custom trained YOLO V5 model is shown in Fig. 7. The loss function of yolo is calculated using the Equation 1:

$$\begin{aligned} \lambda_{coord} \sum_{i=0}^{s^2} \sum_{j=0}^B 1_{ij}^{obj} [(x_i - \hat{x}_i)^2 - (y_i - \hat{y}_i)^2] + \lambda_{coord} \\ \sum_{i=0}^{s^2} \sum_{j=0}^B 1_{ij}^{obj} [(\sqrt{w_i} - \sqrt{\hat{w}_i})^2 - (\sqrt{h_i} - \sqrt{\hat{h}_i})^2] + \\ \sum_{i=0}^{s^2} \sum_{j=0}^B 1_{ij}^{obj} (C_i - \hat{C}_i)^2 + \lambda_{noobj} \sum_{i=0}^{s^2} \sum_{j=0}^B 1_{ij}^{obj} (C_i - \hat{C}_i)^2 \\ + \sum_{i=0}^{s^2} 1_{ij}^{obj} \sum_{c \in \text{cases}} (p_i(c) - \hat{p}_i(c))^2 \end{aligned} \quad (1)$$

YOLO V5 features are flattened and combined with facial landmarks and GLCM features to capture all the nuances of the images. The vast number of features obtained after fusing may lead to overfitting and a significant increase in computation time. Principal Component Analysis (PCA) is applied to reduce the number of features and still preserve the most useful information from the feature set.

The ‘explained_variance’ attribute of PCA class is used to calculate the total variance contributed by the first k principal components based on which the dimensionality reduction is performed.

The heterogeneous ensemble classifier includes KNN, SVM, Random Forest, XGBoost and Multilayer Perceptron (MLP). Performance of the hybrid model is evaluated using the four important measures namely accuracy, precision, recall and F1 score, with their formulae shown in Equations 2-5. With regard to multinomial classification, accuracy gives the proportion of correct predictions out of all the predictions made by the model; Precision gives the proportion of correct predictions out of all the positive predictions made by the model whereas recall indicates the proportion of correct predictions out of all the positive instances of the target class.

$$precision = \frac{TP}{TP + FP} \quad (2)$$

$$recall = \frac{TP}{TP + FN} \quad (3)$$

$$F1 = \frac{2 \times precision \times recall}{precision + recall} \quad (4)$$

$$Accuracy = \frac{TP + TN}{TP + FN + TN + FP} \quad (5)$$

Table 7 summarizes the values of the 4 metrics assessed for the ensemble classifier that makes the final prediction in the proposed hybrid model. To prove the effectiveness of the ensemble model, the five individual classifiers are employed in place of the ensemble model and their performances are also measured and listed in the table. In terms of accuracy, it can be observed that KNN outperforms the other models, when tested for Indian dataset and the combined dataset consisting of images of both Indian and Russian children. However, for the Russian dataset, Random Forest and XGBoost algorithms based on homogeneous ensemble techniques are more accurate than KNN. Apparently, no single model is perfect for all datasets. In contrast, the ensemble model that combines the results of the 5 classifiers consistently performs better than any other individual model, for all types of dataset.

The ensemble model also records the highest recall values among the other models, when evaluated for the 3 types of data. With regard to the precision value, although the ensemble classifier experiences a marginal decrease, when executed for Indian and Russian datasets, still it performs the best for the combined dataset. The superior performance of the ensemble classifier, as shown in Fig. 8a-8d, justifies its inclusion at the final stage of the proposed hybrid model.

While this cross-cultural study demonstrates the potential of the proposed model in emotion classification, further research is required to address a few limitations not covered in this study. Firstly, the limited dataset since the dataset used is a sensitive dataset of audio and video data of children in the age group 5-16. Even though we have

TABLE 7. Performance metrics evaluation with three different datasets for 5 different classifiers.

Dataset	Tamil						Russian						Combined(Russian + Tamil)					
	KNN	SVM	Random Forest	XGBoost	MLP	Ensemble Classifier	KNN	SVM	Random Forest	XGBoost	MLP	Ensemble Classifier	KNN	SVM	Random Forest	XGBoost	MLP	Ensemble Classifier
Accuracy	0.95	0.92	0.92	0.89	0.8684	0.943	0.92	0.81	0.9695	0.9695	0.9254	0.972	0.93	0.86	0.89	0.88	0.9044	0.962
F1 score	0.93	0.9	0.9	0.88	0.86	0.931	0.92	0.79	0.96	0.96	0.9	0.97	0.9	0.84	0.87	0.85	0.88	0.94
Recall	0.93	0.91	0.89	0.87	0.89	0.932	0.91	0.77	0.95	0.95	0.89	0.96	0.91	0.86	0.84	0.83	0.87	0.95
Precision	0.94	0.91	0.92	0.91	0.86	0.93	0.94	0.83	0.98	0.98	0.94	0.97	0.91	0.83	0.92	0.89	0.9	0.93

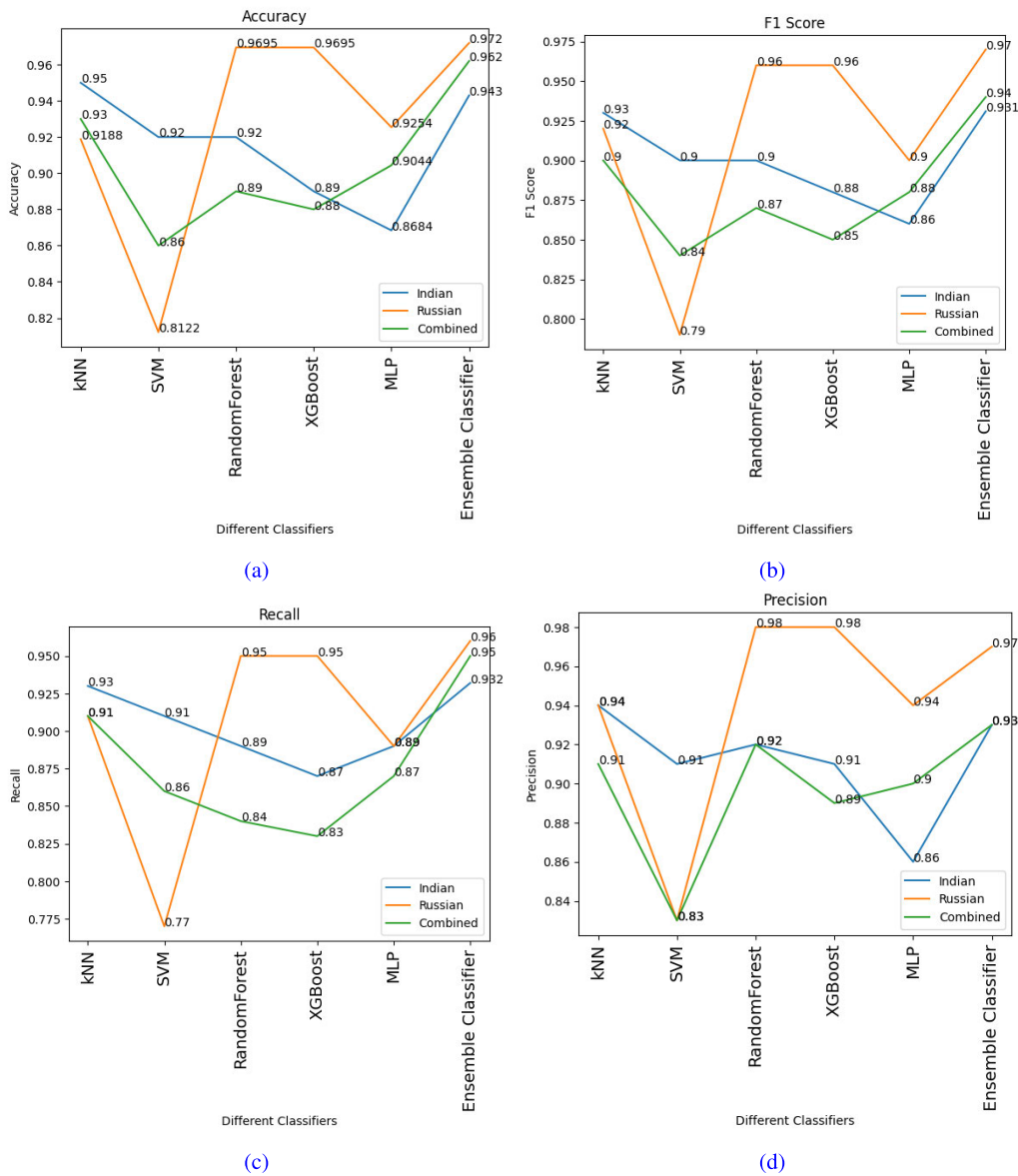


FIGURE 8. Evaluation metrics for KNN, SVM, random forest, XGBoost, MLP and ensemble classifier with three different Datasets- indian, russian, combined. (a) Accuracy (b) Precision (c) Recall (d) F1 Score.

used enough number of frames from each individual data collected from limited children, if the diversity of data is

improved which help to mitigate the problems of skewed classification and overfitting, arising due to imbalanced

TABLE 8. Performance metrics evaluation with different dataset combinations.

Dataset	Models	Performance Metrics			
		Accuracy	F1 Score	Recall	Precision
Russian	YOLO	0.88	0.87	0.88	0.86
	Facial Landmarks	0.65	0.56	0.55	0.66
	GLCM	0.75	0.71	0.69	0.77
	Facial Landmarks + GLCM	0.63	0.63	0.63	0.63
	YOLO +Facial Landmarks + GLCM	0.97	0.97	0.96	0.97
Tamil	YOLO	0.89	0.86	0.86	0.86
	Facial Landmarks	0.75	0.72	0.72	0.76
	GLCM	0.65	0.51	0.54	0.49
	Facial Landmarks + GLCM	0.69	0.68	0.69	0.69
	YOLO +Facial Landmarks + GLCM	0.94	0.93	0.93	0.93
Combined	YOLO	0.87	0.85	0.84	0.88
	Facial Landmarks	0.72	0.69	0.67	0.74
	GLCM	0.64	0.62	0.6	0.69
	Facial Landmarks + GLCM	0.62	0.58	0.62	0.60
	YOLO +Facial Landmarks + GLCM	0.96	0.94	0.95	0.93

and insufficient dataset. Secondly, given the importance of developing interpretable deep learning models to understand the reason behind the predictions for high-stakes domains like healthcare systems, our future research will focus on visualizing and understanding the key input features that influence the decision-making process of the models. Lastly, incorporating optimization techniques could help refine the model and speed up the inference, for deployment in resource-constrained environments, but it is not part of the proposed architecture.

B. ABLATION STUDY

Ablation study is carried out to demonstrate the importance of the various components of the proposed model. Besides the three baseline variants, one using only YOLO features, the second using only Facial landmarks and the third using GLCM features, the experimentation also includes two hybrid model variants. Variant IV integrates the facial landmark and GLCM features. The two features are fused and used for classification by the ensemble model. On the other hand, Variant V is the proposed architecture that includes YOLO V5 in addition to facial landmarks and GLCM features. Table 8 presents the performance metrics of the 5 variants evaluated with the 3 datasets - Tamil, Russian and combined dataset. From Fig.9, it is seen that variant I based exclusively on YOLO features surpasses the results of the other two baseline variants II and III for all datasets. The figure also reveals that, among the variants II and III, variant III based only on GLCM features yields superior results than variant II across all metrics, for Russian data; however, when evaluated for Tamil and combined datasets,



FIGURE 9. Ablation study.

variant II based only on facial landmarks are found to be better than variant III. This observation reinforces the choice

of both the features for developing the hybrid model. Further, the results of the two hybrid variants (Variants IV and V) prove the significance of YOLO V5 features in the proposed architecture. Despite the fact that facial landmarks and GLCM features provide useful information about the texture and facial expressions in the image, the performance metrics with only those features being less than 70% in Table 8, shows that these features do not capture all the nuances of the image, essential for the recognition of facial expressions. On the contrary, variant V exhibits a drastic improvement in the performance with the inclusion of YOLO V5, as depicted in Fig. 9.

V. CONCLUSION

A hybrid model is developed to facilitate efficient cross-cultural research on emotion recognition from facial images of Indian and Russian children. The hybrid model is designed by integrating facial landmark features, GLCM features and YOLO V5 extracted features, with a thorough literature analysis proving the effectiveness of these features. The custom cross-cultural dataset for this study is built with the facial images of 28 Indian Tamil-speaking children (10 boys and 18 girls) and 64 Russian children (32 boys and 32 girls), in the age group of 5 to 16 years, portraying the four emotions - happy, angry, neutral and sad. A majority-voting heterogeneous ensemble classifier, that includes KNN, SVM, XGBoost, Random Forest and Multilayer Perceptron, is employed at the final stage of the hybrid model to make the multinomial classification. Feature reduction is done to balance the increased number of features after fusion.

Upon evaluating the model with three groups of datasets - Indian, Russian and combined dataset comprising images of both Indian and Russian children, it is seen that the ensemble classifier outperforms its individual counterparts. An ablation study carried out to assess the contribution of the various components of the hybrid model proves the phenomenal contribution of YOLO V5 extracted features, compared to GLCM and facial landmark features. Based on the ensemble classifier results, it shows that the separate Russian dataset slightly performs well then the combined dataset, this is obvious that separate datasets implicitly have at least a domain shift related to the cultural differences. But from the results it is evident that, the model can be utilized for the emotional recognition of dataset with children's facial data with different cultural background. Nevertheless, empirical results support the established literature on the positive impact of feature fusion in improving the classification accuracy, with facial landmarks and GLCM providing useful information about the facial expressions and texture of the image. The capability of YOLO V5 to detect small objects and work with low-light images justifies the choice of YOLO V5 to build the hybrid model. The proposed model can be used in real time as YOLO V5 is a single-stage object detector and thus is super-fast. Cross-cultural study

has helped to reduce bias in the formulation of emotion recognition model.

As a future work, datasets from other regions can be included. Multinomial classification can be extended to a greater number of emotions and various other deep neural networks can be explored for the same. The study reveals promising results for early detection of emotional disorders and delayed development of emotional manifestation in typically developing children. This study can be extended for analysing emotional manifestations of atypically developed children.

ACKNOWLEDGMENT

The authors would like to thank all the children who participated in this research work and their parents/caretakers.

They also like to thank Vellore Institute of Technology, Vellore, for providing necessary technical support and the resources to carry out this research.

REFERENCES

- [1] M. Roshan, M. Rawat, K. Aryan, E. Lyakso, A. M. Mekala, and N. Ruban, "Linguistic based emotion analysis using softmax over time attention mechanism," *PLoS ONE*, vol. 19, no. 4, Apr. 2024, Art. no. e0301336.
- [2] A. R. Khan, "Facial emotion recognition using conventional machine learning and deep learning methods: Current achievements, analysis and remaining challenges," *Information*, vol. 13, no. 6, p. 268, May 2022.
- [3] E. Lyakso, N. Ruban, O. Frolova, and M. A. Mekala, "The children's emotional speech recognition by adults: Cross-cultural study on Russian and Tamil language," *PLoS ONE*, vol. 18, no. 2, Feb. 2023, Art. no. e0272837.
- [4] R. Nersissson, P. Bhuyan, A. M. Mekala, and E. Lyakso, "Automatic emotion recognition system: A cross culture study between Tamil and Russian speaking children," in *Proc. 3rd Int. Conf. Adv. Res. Comput. (ICARC)*, Feb. 2023, pp. 84–89.
- [5] E. Lyakso, O. Frolova, N. Ruban, and A. M. Mekala, "Child's emotional speech classification by human across two languages: Russian & Tamil," in *Proc. 23rd Int. Conf. Speech Comput.*, St. Petersburg, Russia, Cham, Switzerland: Springer, 2021, pp. 384–396.
- [6] E. Lyakso, O. Frolova, A. Nikolaev, E. A. Kleshnev, P. Grave, A. Ilyas, O. Makhnytkina, R. Nersissson, A. M. Mekala, and M. Varalakshmi, "Recognition of the emotional state of children by video and audio modalities by Indian and Russian experts," in *Proc. Int. Conf. Speech Comput.*, Jan. 2023, pp. 469–482.
- [7] M. Kumar, N. Katyal, N. Ruban, E. Lyakso, A. Mary Mekala, A. N. Joseph Raj, and G. Maarc Richard, "Transfer learning based convolution neural net for authentication and classification of emotions from natural and stimulated speech signals," *J. Intell. Fuzzy Syst.*, vol. 41, no. 1, pp. 2013–2024, Aug. 2021.
- [8] S. Tariyal, R. Chauhan, Y. Bijalwan, R. Rawat, and R. Gupta, "A comparative study of MTCNN, viola-jones, SSD and YOLO face detection algorithms," in *Proc. Int. Conf. Intell. Innov. Technol. Comput., Electr. Electron. (IITCEE)*, Jan. 2024, pp. 1–7.
- [9] K. Anusudha, "Real time face recognition system based on YOLO and InsightFace," *Multimedia Tools Appl.*, vol. 83, no. 11, pp. 31893–31910, Sep. 2023.
- [10] H. Lv, H. Yan, K. Liu, Z. Zhou, and J. Jing, "YOLOv5-AC: Attention mechanism-based lightweight YOLOv5 for track pedestrian detection," *Sensors*, vol. 22, no. 15, p. 5903, Aug. 2022.
- [11] H. Ouanan, M. Ouanan, and B. Aksasse, "Facial landmark localization: Past, present and future," in *Proc. 4th IEEE Int. Colloq. Inf. Sci. Technol.*, Oct. 2016, pp. 487–493.
- [12] F. Khan, "Facial expression recognition using facial landmark detection and feature extraction via neural networks," 2018, *arXiv:1812.04510*.

- [13] S. A. Alazawi, N. M. Shati, and A. H. Abbas, "Texture features extraction based on GLCM for face retrieval system," *Periodicals Eng. Natural Sci. (PEN)*, vol. 7, no. 3, p. 1459, Oct. 2019.
- [14] B. Niu, Z. Gao, and B. Guo, "Facial expression recognition with LBP and ORB features," *Comput. Intell. Neurosci.*, vol. 2021, no. 1, pp. 1–10, Jan. 2021.
- [15] F. H. Almkhtar, "Facial emotions recognition using local monotonic pattern and gray level co-occurrence matrices plant leaf images aided agriculture development," *Optik*, vol. 271, Dec. 2022, Art. no. 170161.
- [16] D. Mehta, M. F. H. Siddiqui, and A. Y. Javaid, "Recognition of emotion intensities using machine learning algorithms: A comparative study," *Sensors*, vol. 19, no. 8, p. 1897, Apr. 2019.
- [17] N. Raut, "Facial emotion recognition using machine learning," San Jose State Univ., San Jose, CA, USA, Master's Projects 632, 2018, doi: [10.31979/etd.w5fs-s8wd](https://doi.org/10.31979/etd.w5fs-s8wd).
- [18] A. I. Siam, N. F. Soliman, A. D. Algarni, F. E. Abd El-Samie, and A. Sedik, "Deploying machine learning techniques for human emotion detection," *Comput. Intell. Neurosci.*, vol. 2022, pp. 1–16, Feb. 2022.
- [19] G. Meena, K. K. Mohbey, A. Indian, M. Z. Khan, and S. Kumar, "Identifying emotions from facial expressions using a deep convolutional neural network-based approach," *Multimedia Tools Appl.*, vol. 83, no. 6, pp. 15711–15732, Jul. 2023.
- [20] J. H. Kim, A. Poullose, and D. S. Han, "CVGG-19: Customized visual geometry group deep learning architecture for facial emotion recognition," *IEEE Access*, vol. 12, pp. 41557–41578, 2024.
- [21] J. H. Kim, A. Poullose, and D. S. Han, "The extensive usage of the facial image thresholding machine for facial emotion recognition performance," *Sensors*, vol. 21, no. 6, p. 2026, Mar. 2021.
- [22] A. S. A. Hans and S. Rao, "A CNN-LSTM based deep neural networks for facial emotion detection in videos," *Int. J. Adv. Signal Image Sci.*, vol. 7, no. 1, pp. 11–20, Mar. 2021.
- [23] H. V. Manalu and A. P. Rifai, "Detection of human emotions through facial expressions using hybrid convolutional neural network-recurrent neural network algorithm," *Intell. Syst. Appl.*, vol. 21, Mar. 2024, Art. no. 200339.
- [24] Z. Ullah, L. Qi, A. Hasan, and M. Asim, "Improved deep CNN-based two stream super resolution and hybrid deep model-based facial emotion recognition," *Eng. Appl. Artif. Intell.*, vol. 116, Nov. 2022, Art. no. 105486.
- [25] A. J. Obaid and H. K. Alrammahi, "An intelligent facial expression recognition system using a hybrid deep convolutional neural network for multimedia applications," *Appl. Sci.*, vol. 13, no. 21, p. 12049, Nov. 2023.
- [26] B. Assiri and M. A. Hossain, "Face emotion recognition based on infrared thermal imagery by applying machine learning and parallelism," *Math. Biosci. Eng.*, vol. 20, no. 1, pp. 913–929, 2022.
- [27] C. Kyal, H. Poddar, and M. Reza, "Human emotion recognition from spontaneous thermal image sequence using GPU accelerated emotion landmark localization and parallel deep emotion net," in *Proc. Int. Conf. Innov. Comput. Commun.*, vol. 1, Aug. 2020, pp. 931–943.
- [28] Z. Song, "Facial expression emotion recognition model integrating philosophy and machine learning theory," *Frontiers Psychol.*, vol. 12, Sep. 2021, Art. no. 759485.
- [29] T. Debnath, M. M. Reza, A. Rahman, A. Beheshti, S. S. Band, and H. Alinejad-Rokny, "Four-layer ConvNet to facial emotion recognition with minimal epochs and the significance of data diversity," *Sci. Rep.*, vol. 12, no. 1, p. 6991, Apr. 2022.
- [30] R. Nersisson, T. J. Iyer, A. N. Joseph Raj, and V. Rajangam, "A dermoscopic skin lesion classification technique using YOLO-CNN and traditional feature model," *Arabian J. Sci. Eng.*, vol. 46, no. 10, pp. 9797–9808, Oct. 2021.
- [31] T. J. Iyer, K. Rahul, R. Nersisson, Z. Zhuang, A. N. J. Raj, and I. Refayee, "Machine learning-based facial beauty prediction and analysis of frontal facial images using facial landmarks and traditional image descriptors," *Comput. Intell. Neurosci.*, vol. 2021, no. 1, Jan. 2021, Art. no. 4423407.
- [32] E. Lyakso, O. Frolova, E. Kleshnev, N. Ruban, A. M. Mekala, and K. V. Arulalan, "Approbation of the child's emotional development method (CEDM)," in *Proc. Int. Conf. Multimodal Interact.*, Nov. 2022, pp. 201–210.
- [33] D. T. Cordaro, R. Sun, D. Keltner, S. Kamble, N. Huddar, and G. McNeil, "Universals and cultural variations in 22 emotional expressions across five cultures," *Emotion*, vol. 18, no. 1, pp. 75–93, Feb. 2018.
- [34] S. Fridenson-Hayo, S. Berggren, A. Lassalle, S. Tal, D. Pigat, S. Bölte, S. Baron-Cohen, and O. Golan, "Basic and complex emotion recognition in children with autism: Cross-cultural findings," *Mol. Autism*, vol. 7, no. 1, pp. 1–11, Dec. 2016.
- [35] Y. Huang, S. Tang, D. Helmeeste, T. Shioiri, and T. Someya, "Differential judgement of static facial expressions of emotions in three cultures," *Psychiatry Clin. Neurosci.*, vol. 55, no. 5, pp. 479–483, Oct. 2001.
- [36] E. Lyakso, O. Frolova, A. Nikolaev, S. Grechanyi, A. Matveev, Y. Matveev, O. Makhnytkina, and R. Nersisson, "Emotional state of children with ASD and intellectual disabilities: Perceptual experiment and automatic recognition by video, audio and text modalities," in *Proc. Int. Conf. Speech Comput.*, Jan. 2023, pp. 535–549.



A. MARY MEKALA was born in India, in 1984. She received the bachelor's degree in computer science and engineering from Anna University, India, in 2006, the master's degree in computer science from Sathyabama University, India, in 2008, and the Ph.D. degree in computer science from Vellore Institute of Technology, Vellore, India, in 2019.

She has more than 15 years of teaching experience and is currently an Associate Professor with Vellore Institute of Technology. Her research interests include machine learning and network security. She has published more than ten papers in Scopus indexed journals and conferences related to these fields. She also undertakes and executes consultancy projects for startups.



M. VARALAKSHMI was born in India, in 1978. She received the bachelor's degree in computer science and engineering from Madras University, India, in 2000, and the master's (Hons.) and Ph.D. degrees in computer science from Vellore Institute of Technology (VIT), Vellore, India, in 2013 and 2019, respectively.

She has more than 18 years of academic experience and is currently an Associate Professor with VIT. Her research interests include natural language processing, deep learning, and high-performance computing. She is particularly interested in image classification algorithms and large language models. She has published more than ten papers in SCI and Scopus indexed journals and reputed national/international conferences related to these fields. She completed a project funded by ISRO and is currently working on an Indo-Russian joint research project, funded by DST. She also works on consultancy projects for startups.



C. P. ACHYUTHA GOWDA received the B.E. degree in electronics and communication engineering from the BMS Institute of Technology and Management (affiliated to Visvesvaraya Technological University), Bengaluru, Karnataka, India, in 2024.

He gained research experience as an Intern at the Department of Science and Technology (DST), working under Dr. N. Ruban at Vellore Institute of Technology.



LETI MANISH KUMAR was born in Andhra Pradesh, India, in 2002. He received the B.E. degree in electronics and communication engineering from the BMS Institute of Technology and Management (affiliated to Visvesvaraya Technological University), Bengaluru, Karnataka, India, in 2024.

He gained research experience as an Intern at the Department of Science and Technology (DST), working under Dr. N. Ruban at Vellore Institute of Technology.



ELENA E. LYAKSO received the Ph.D. degree in neuroscience and the Dr.-Sc. degree in speech psychophysiology from Saint Petersburg State University.

She is currently the Head of the Laboratory Child Speech Research Group, Department of Higher Nervous Activity and Psychophysiology, Biological Faculty, Saint Petersburg State University, Russia. She has published five books, four textbooks, and more than

300 scientific articles on physiological and neurological factors influenced on language acquisition, emotional sphere of children, biological and physiological basis of child speech development in ontogenesis and dysontogenesis. She is the author of original lecture courses and practices for bachelor's, master's, and Ph.D. students of biology, psychology and philology faculties at Saint Petersburg State University. She supervises more than 30 bachelor's, master's, Ph.D. students, and post-docs. She was the supervisor of international projects with Turku University, Finland; Helsinki University, Finland; and University of Amsterdam, The Netherlands; and projects supported by Russian Government and Russian Foundation for Basic Research. Her current research projects supported by Russian Science Foundation and international projects with Vellore Institute of Technology, India, are aimed is to develop a cross-cultural approach for the diagnosis and future correction of emotional disorders in children, taking into account the age, cultural and linguistic affiliation of children and the severity of psychoneurological and developmental disorders; and recognize the emotional states of children in voice, speech, facial expressions using artificial intelligence methods.

Dr. Lyakso is a member of European Psychology Society, Saint Petersburg Society of Naturalists, Acoustical Society, Physiological Society, and the Head of the Section on Speech Physiology at Russian Physiological Society.



OLGA FROLOVA was born in Russia, in 1979. She received the M.S. degree in biology and the Ph.D. degree in psychophysiology from Saint Petersburg State University, in 2002 and 2008, respectively.

She is currently a Researcher with the Laboratory Child Speech Research Group, Department of Higher Nervous Activity and Psychophysiology, Biological Faculty, Saint Petersburg State University, Russia. She is the author of more than 60 publications on the speech development of

typically developing children, children with intellectual disabilities and mixed specific developmental disorders, the head of the grants supported by Russian Foundation for Basic Research and performer of grant supported by Russian Science Foundation. Her research interests include the speech development of orphans, vocal-speech interaction between a mother and a child, and acoustic characteristics of speech of children with intellectual disabilities.



RUBAN NERSISSON was born in India, in 1980. He received the Bachelor of Engineering degree in instrumentation and control engineering from the University of Madras, India, in 2001, the master's degree in biomedical signal processing and instrumentation from SASTRA University, Thanjavur, India, in 2004, and the Ph.D. degree from VIT University, India, in 2018.

He has more than 20 years of teaching experience in various engineering colleges in and out of India. He is currently an Academic Faculty Member with VIT University. His research interests include bio signal processing, speech recognition, and machine learning. He has more than 70 research publications in SCI, Scopus indexed journal, and national/international reputed conferences on the above mentioned fields. He has six book chapters and two research hand book in his credit.

...