Discover Artificial Intelligence

Research

FER-BHARAT: a lightweight deep learning network for efficient unimodal facial emotion recognition in Indian context

Ruhina Karani[1] · **Jay Jani**[2] · **Sharmishta Desai**[1]

Received: 24 July 2023 / Accepted: 11 May 2024
Published online: 15 May 2024

Abstract
Humans' ability to manage their emotions has a big impact on their ability to plan and make decisions. In order to better understand people and improve human–machine interaction, researchers in affective computing and artificial intelligence are investigating the detection and recognition of emotions. However, different cultures have distinct ways of expressing emotions, and the existing emotion recognition datasets and models may not effectively capture the nuances of the Indian population. To address this gap, this study proposes custom-built lightweight Convolutional Neural Network (CNN) models that are optimized for accuracy and computational efficiency. These models are trained and evaluated on two Indian emotion datasets: The Indian Spontaneous Expression Dataset (ISED) and the Indian Semi Acted Facial Expression Database (iSAFE). The proposed CNN model with manual feature extraction provides remarkable accuracy improvement of 11.14% for ISED and 4.72% for iSAFE datasets as compared to baseline, while reducing the training time. The proposed model also surpasses the accuracy produced by pre-trained ResNet-50 model by 0.27% ISED and by 0.24% for the iSAFE dataset with significant improvement in training time of approximately 320 s for ISED and 60 s for iSAFE dataset. The suggested lightweight CNN model with manual feature extraction offers the advantage of being computationally efficient and more accurate compared to pre-trained model making it a more practical and efficient solution for emotion recognition among Indians.

Keywords Human emotion recognition · Affective computing · Convolution neural network · Indian emotion · Facial expressions

1 Introduction

Artificial intelligence (AI) has replaced conventional approaches in a variety of disciplines of interest in today's digitally-driven world [1, 2]. The world is quickly embracing AI-based methods in industries including social media, banking, healthcare, and education. A precise identification of human emotion is a necessity for some AI-based applications, including patient care in the medical industry, understanding level measurement in the educational domain, and identifying emotional intelligence in psychology, among others [3]. The field of human emotion recognition has seen a lot of study, but very little of it has focused on data samples with Indian origins. Although, some literature is available on identifying human emotions using machine learning and deep learning based approach, most of it concentrates on the data available from European and American continents.

✉ Ruhina Karani, ruhina.karani@djsce.ac.in; Jay Jani, jayjani482001@gmail.com; Sharmishta Desai, sharmishta.desai@mitwpu.edu.in | [1]School of Computer Engineering and Technology, Dr. Vishwanath Karad MIT World Peace University, Pune, India. [2]The Ohio State University, Columbus, OH, USA.

Additionally, the majority of studies use a computer vision-based techniques to determine emotions. India, being a nation with a diversified history and rich cultural legacy, has different ways of expressing emotions than the rest of the globe [3]. Indians are frequently observed to have less expressive facial expressions when expressing their emotions. As a result, it's possible that AI-based machine learning and deep learning models trained on a single ethnicity dataset won't deliver trustworthy findings when identifying Indians' emotions precisely. Also, the existing pre trained deep learning models are computationally complex and require large training time for unimodal emotion recognition.

To address these gaps, this paper proposes custom-made lightweight CNN models for efficient unimodal emotion identification among Indians, with reduced computational complexity. Due to the limited availability of multimodal data for Indian Emotion Samples, this work adheres to a unimodal approach of emotion recognition utilizing facial expression images.

The research aims to apply these models on the unimodal datasets having video emotion samples collected from the people of India in order to improve the baseline results in terms of accuracy with reduced training time. In order to achieve this, the paper concentrates of two datasets ISED [4] and iSAFE [5] since both of these datasets contain the emotion samples of Indians in the age group of 17–22 years. Happy et al. [4] in their publication of the dataset "The India Spontaneous Expression Database (ISED)" for Emotion Recognition, reported the highest accuracy as 86.46% by applying Local Gabor Binary Pattern (LGBP) operator for feature extraction and PCA with Linear Discriminant Analysis (LDA) for classification and set it as a baseline for future reference. In the publication, "Indian Semi-Acted Facial Expression (iSAFE) for Human Emotions Recognition," Singh, S. and Benedict, Shajulin [5] achieved 92% accuracy by applying the Convolution Neural Network (CNN) ResNet34 model for classification and established it as a benchmark for subsequent comparison. This study proposes to extract important features from facial images and use custom built lightweight CNN architecture in order to achieve improved accuracy.

The paper is organized as follows: Sect. 2 describes a thorough literature review. The topic of the suggested methodology is covered in Sect. 3. The findings and discussions are described in Sect. 4. The paper is concluded in Sect. 5.

2 Related work

The related work is divided into two sections. Section A describes all the datasets explored with the intent of discovering datasets having emotions of Indians. Section B describes the review of literature carried out for machine learning and deep learning techniques used for identifying human emotions.

2.1 Existing datasets for emotion recognition

This section focuses on different datasets available for human emotion recognition. The review was conducted to explore different state of art datasets on human emotions and finalize the datasets based on availability of samples of Indian population. The datasets available are bifurcated based on the modality in which the data is available. Some datasets have the data of single modality like images or audio and some datasets provide multimodal data from audio visual modality, EEG signal etc.

Mollahosseini Ali et al. [6] described the single-modality dataset Affectnet, which contains pictures of people's facial expressions along with the valence and arousal levels of those emotions. The existence of seven distinct face emotions (categorical model) was manually noted in around half of the returned photos (440 K), including Happy, Sad, Surprise, Fear, Disgust, Anger, and Contempt.

The ASCERTAIN dataset, which includes data from many modalities including electroencephalogram (EEG), electrocardiogram (ECG), galvanised skin response (GSR), and facial activity, was proposed by M. R. Subramanian et al. [7] The collection includes data on 58 individuals with a mean age of 30 years (37 men and 21 women). Lucey et al. [8] suggested an Extended Cohn-Kanade (CK+) video dataset of 593 video sequences from 123 distinct people, ranging in age from 18 to 50, with a range of gender identities and ethnic backgrounds. One of the seven expression classes—anger, contempt, disgust, fear, pleasure, sorrow, and surprise—is assigned to 327 of these movies.

Kosti et al. [9] offered a single modality EMOTIC image dataset with 23, 571 images and dimensions for valence, arousal, and dominance. Peace, Affection, Esteem, Anticipation, Engagement, Confidence, Happiness, Pleasure, Excitement, Surprise, Sympathy, Doubt/Confusion, Disconnection, Fatigue, Embarrassment, Yearning, Disapproval, Aversion, Annoyance, Anger, Sensitivity, Sadness, Disquietment, Fear, Pain and Suffering are some of the 26 emotion categories annotated on the images. 3500 labelled images make up the development set, 3,500 labelled images make up the test set, and 28,000

labelled images make up the training set of the proposed image dataset FER-2013 [10]. The seven emotions happy, sad, angry, terrified, surprised, disgusted are assigned to each image in FER-2013.

The Google Facial Expression Comparison Dataset [11] is a dataset of facial expression images of a single modality that consists of triplets of faces with human annotations indicating which two faces in each triplet have the most comparable facial expressions. Six or more human raters each annotated one triplet in this sample. 500–156 K triplets and 156 K face photos make up the dataset. Park et al. [12] introduced K-EmoCon, a multimodal sensor dataset for continuous emotion recognition in naturalistic conversations that includes peripheral physiological signals, audiovisual recordings, and EEG. The multimodal audio visual dataset eNTERFACE'05 introduced by O. Martin et al. [13] comprises of 1166 video sequences with the six emotions of happiness, sadness, surprise, fear, disgust, and anger. The dataset's participants come from a variety of countries, including Belgium, Cuba, Turkey, Slovakia, France, Brazil, Spain, the United States, Greece, Croatia, Italy, Canada, Austria, and Russia.

Busso et al. [14] at the sail lab at USC developed the multimodal and multilingual database known as IEMOCAP. Two hours' worth of audiovisual material, comprising video, speech, face motion capture, and text transcriptions of 10 actors—5 men and 5 women—are included in the database. It comprises of dyadic sessions in which actors act out improvised scenes or scripted situations that have been chosen intentionally to elicit emotional reactions. Multiple annotators have added category labels to the IEMOCAP database, such as "angry," "happy," "sad," and "neutral," as well as dimensional labels like "valence," "activation," and "dominance." JAFFE by Michael J. Lyons et al. [15, 16] is a single modality image dataset of face expression that includes 213 photographs of 7 different facial expressions that were modelled by 10 Japanese women. Soujanya Poria et al. [17] introduced MELD; a Multimodal Audio Visual Emotion Dataset that includes text, audio, and visual modalities for dialogue occurrences. More than 1400 lines and 13000 utterances from the Friends TV series are available in MELD. A dialogue's words have been assigned the emotions of Anger, Disgust, Sadness, Joy, Neutral, Surprise, and Fear. A multimodal library of speech and song called RAVDESS; introduced by Livingstone SR and Russo FA [18] has the recordings of 24 professional actors who perform lexically-matched phrases in a neutral North American accent. Speech comprises expressions that are peaceful, pleased, sad, angry, afraid, surprised, and disgusted.

SEWA, a multimodal dataset by Kossaifi et al. [19] is a collection of annotated audio and 2D visual dynamic behavior. Features recordings of six volunteer groups, each with 30 participants, representing six distinct ethnic groups: British, German, Hungarian, Greek, Serbian, and Chinese. The final database includes recordings from 199 experiment sessions, including more than 550 min of recorded computer-mediated face-to-face interactions between subject pairs and 1525 min of audio-visual data on people's responses to advertisements from 398 different subjects. The Indian Spontaneous Expression Database (ISED) [4] is a single-modality video dataset of Indian-natural face expressions. The dataset contains 428 video clips of 50 healthy volunteers, including 29 men and 21 women, who are between the ages of 18 and 22 and come from various parts of India. Four emotions—happiness, surprise, sadness, and disgust—were noted in the samples.

The Center for Intelligent Sensing created the single modality Image Dataset of Indian Facial Expression known as IFExD [20]. 342 photos of two sets of faces taken from the "Front," "Left," and "Right" perspectives make up the dataset. Eight distinct facial expressions have been recorded for each angle of the face. "Happy," "Sad," "Neutral," "Disgust," "Fear," "Surprise," "Contempt," and "Anger" are among them. A single modality video collection of Indian semi-acted facial expression is called iSAFE [5]. The dataset includes 395 clips of 44 volunteers between the ages of 17 and 22; face expressions were recorded as they watched a few stimulant videos; volunteers self-annotated the facial expressions, which were then cross-annotated by annotators for the categories of happiness, sadness, surprise, disgust, fear, anger, uncertainty, and no emotion. Arunashri et al. [21] introduced the IFED dataset, which includes 112 participant films for the basic seven emotions of anger, contempt, disgust, fear, happiness, sorrow, and surprise, is a single-modality video dataset of Indian facial expressions.

A multimodal Video Emotion Dataset called HEU Emotion was introduced by Chen Jing et al. [22] containing 19,004 video clips split into two halves. Ten emotions and two modalities are depicted in videos that were downloaded from Tumblr, Google, and Giphy in the first section (facial expression and body posture). The second section contains a corpus composed of 10 emotions and 3 modalities that was carefully collected from movies, TV shows, and variety shows (facial expression, body posture, and emotional speech). The Toronto Emotional Speech Collection (TESS) published by Pichora Fuller et al. [23], is a single modality Audio Emotion Dataset in which two actresses (aged 26 and 64) express each of the seven emotions by speaking a set of 200 target words in the carrier phrase "Say the word." anger, disgust, fear, happiness, pleasant surprise, sadness, and neutral. The single modality Audio Emotion Dataset ASVP-ESD (Speech & Non-Speech Emotional Sound) by Dejoli et al. [24], has 5146 audio files (with additional 1204 files for baby's voices). It is a database that is based on emotions and contains both speech and non-spoken emotional sound. The audio was

recorded and gathered from emotional sound websites, TV shows, movies, and YouTube channels. SAVEE [25] is a multimodal Audio Visual Dataset that includes 480 British English utterances recorded from four male actors expressing the six fundamental emotions and neutral emotions. The words were selected from the common TIMIT corpus. A single modality Audio Emotion Dataset called the Emotional Speech Database (ESD) introduced by Kun Zhou et al. [26] has 350 parallel utterances made by 10 native Mandarin speakers and 10 native English speakers in five different emotional states neutral, happy, angry, sad and surprise.

Table 1 describes summary of available datasets in terms of modality, emotion categories, number of participants, the presence of Indian emotion samples and No. of citations. The number of citations available for each dataset is extracted from the sources like Google scholar, Kaggle etc.

Figure 1 describes taxonomy of surveyed databases. The datasets are bifurcated into unimodal and multimodal datasets in stage 1 and further segregated into image datasets, video datasets and audio datasets for unimodal and audio visual & other modalities for multimodal category.

2.2 Review of literature for emotion identification

Audiovisual signals can be used to discern between different human emotions. However, it is commonly acknowledged that Indians differ from the rest of the world when expressing oneself through facial expressions [3]. A lot of work is carried out in the field of emotion recognition with single modality like facial expressions, audio or multimodal signals together like audio visual signals with EEG [2]. Happy et al. [4] applied Principal Component Analysis (PCA) with Linear Discriminant Analysis (LDA) on the ISED dataset after extracting features using the Local Gabor Binary Pattern (LGBP), for emotion recognition of Indians and achieved 86.46% accuracy. Using the CNN ResNet34 model on the iSAFE dataset, Shivendra Singh and Shajulin Benedict [5] classified Indians' facial expressions with 92% accuracy.

Using audiovisual signals, Liam Schoneveld et al. [27] created a multimodal emotion identification system. For facial expression identification, the authors suggested a deep CNN model trained with knowledge distillation, and for auditory expression recognition, they suggested a modified and improved VGGish model. For the purpose of recognizing facial emotions, the scientists employed the AffectNet and Google Facial Expression Comparison (FEC) datasets. An audio-visual fusion model combining deep learning features with a Mixture of Brain Emotional Learning (MoBEL) model inspired by the limbic system of the brain was proposed by Zainab Farhoudi et al. [28]. Convolutional neural networks (CNN) and recurrent neural networks (RNN) were utilised by the authors as deep learning techniques to represent highly abstract features, and a fusion model called MoBEL was used to simultaneously learn the previously combined audio-visual information. On the eNterface'05 audio visual, the authors trained and assessed the suggested model.

Babajee et al. [29, 30] proposed a CNN model to recognise facial expressions with an accuracy of 79.8%. The authors used the (FER2013) dataset to detect emotions. Lee et al. [31] introduced the improved BERT model for emotion recognition by combining it with heterogeneous characteristics based on language, auditory, and visual modalities. The authors evaluated the widely used CMU-MOSI, CMU-MOSEI, and IEMOCAP datasets for multimodal sentiment analysis. Darapaneni et al. [32] proposed a deep learning-based method for identifying emotions in a person using their facial expressions in order to ascertain that person's mental state. The researchers employed a deep CNN VGG 16 model and transfer learning techniques on the JAFFE dataset to achieve test accuracy of roughly 88%. After extracting features using Histogram of Oriented Gradients (HOG), Roy Supta et al. [33] implemented support vector machine (SVM) on JAFFE and Cohn-Kanade databases and obtained 97.62% and 98.61% accuracy respectively.

Zhang et al. [34] presented multimodal emotion recognition using speech, video, and text. The team used DenseNet, a convolutional neural network on the IEMOCAP dataset, and obtained 68.38% accuracy in extracting the emotional elements of the video. Nemati [35] developed hybrid fusion for emotion recognition using auditory, visual, and textual modalities. Results of feature-level canonical correlation analysis (CCA) on audio and visual modalities were combined with user comments using a decision-level fusion. The authors tested their approaches using SVM and the Naive Bayes algorithm on the DEAP dataset.

Pablo Barros [36] in their research published a lightweight deep neural network for facial expression recognition having 10 convolution layers and classified the facial expressions into valence and arousal.

From the literature review conducted and summarized, it is evident that even though a lot of research has been done on the subject of emotion recognition, majority of the research conducted does not focus on understanding the emotions of Indian population. The literature reviewed for datasets depicts that there are four datasets having Indian samples out of which citation details are not available for IFExD [20] and IFED [21] datasets. The citation details and references available for the remaining two Indian datasets [4, 5] indicate that not much work is done for

Table 1 Concise summary of datasets reviewed

Dataset	Unimodal	Multimodal	Modalities Covered	Emotion Categories	No. Of Samples	No. Of Participants	Samples Of Indians	No. Of Citations
Affectnet [6]	✓	✗	Images	Happy, Sad, Surprise, Fear, Disgust, Anger And Contempt	1 M	N/A	✗	1059
Ascertain [7]	✗	✓	EEG, ECG, GSR And Facial Activity	N/A	36 Video Clips	58	✗	331
Extended Cohn-Kanade (CK+) [8]	✓	✗	Video	Anger, Contempt, Disgust, Fear, Happiness, Sadness, And Surprise	593 Video Sequences	123	✗	4091
EMOTIC [9]	✓	✗	Images	Peace, Affection, Esteem, Anticipation, Engagement, Confidence, Happiness, Pleasure, Excitement, Surprise, Sympathy, Doubt/Confusion, Disconnection, Fatigue, Embarrassment, Yearning, Disapproval, Aversion, Annoyance, Anger, Sensitivity, Sadness, Disquietment, Fear, Pain And Suffering	23,571 Images	34, 320	✗	34
FER-2013 [10]	✓	✗	Images	Happy, Sad, Angry, Afraid, Surprise, Disgust, And Neutral	28,000 Images	N/A	✗	85
Google Facial Expression Comparison Dataset [11]	✓	✗	Images	Amusement, Anger, Awe, Boredom, Concentration, Confusion, Contemplation, Contempt, Contentment, Desire, Disappointment, Disgust, Distress, Doubt, Ecstasy, Elation, Embarrassment, Fear, Interest, Love, Neutral, Pain, Pride, Realization, Relief, Sadness, Shame, Surprise, Sympathy, Triumph	500 K Triplets And 156 K Face Images	Flickr	✗	27

Table 1 (continued)

Dataset	Unimodal	Multimodal	Modalities Covered	Emotion Categories	No. Of Samples	No. Of Participants	Samples Of Indians	No. Of Citations
K-Emocon [12]	✗	✓	Audio Visual Recordings, EEG, And Peripheral Physiological Signals	Arousal/Valence, Cheerful/Happy/Angry/Nervous/Sad, Boredom/Confusion/Delight/Engaged Concentration/Frustration/Surprise/None, Confrustion/Contempt/Dejection/Disgust/Eureka/Pride/Sorrow/None	172.92 Min Footage	32 Participants In 16 Paired Debates	✗	59
Enterface'05 [13]	✗	✓	Audio And Video	Happy, Sad, Surprise, Fear, Disgust And Anger	1166 Video Sequences	42 From 14 Different Nationalities	✗	639
IEMOCAP [14]	✗	✓	Audio Visual Data, Including Video, Speech, Motion Capture Of Face, Text Transcriptions	Anger, Happiness, Sadness, Neutrality	2 Hours Of Audio Visual Data	10 Actors	✗	2253
JAFFE [15, 16]	✓	✗	Images	6 Basic Facial Expressions And Neutral	213 Images	10 Actors	✗	49
MELD [17]	✗	✓	Audio And Video	Anger, Disgust, Sadness, Joy, Neutral, Surprise And Fear	More Than 1400 Dialogues And 13000 Utterances	Friends TV Series	✗	441
RAVDESS [18]	✗	✓	Speech And Song	Includes Calm, Happy, Sad, Angry, Fearful, Surprise, And Disgust	7356 Recordings	24 Professional Actors	✗	1022
SEWA [19]	✗	✓	Audio And Video	Arousal, Valence And Liking / Disliking, Head Gesture (Nod / Shake) And Facial Action Unit (FAU)	199 Sessions Of Experiment Recordings	6 Groups Of Volunteers (30 Persons Per Group	✗	144
ISED [4]	✓	✗	Videos	Happiness, Surprise, Sadness And Disgust	428 Video Clips	50	✓	87
Ifexd [20]	✓	✗	Images	Happy, Sad, Neutral, Disgust, Fear, Surprise, Contempt And Anger	342 Images	N/A	✓	N/A

Table 1 (continued)

Dataset	Unimodal	Multimodal	Modalities Covered	Emotion Categories	No. Of Samples	No. Of Participants	Samples Of Indians	No. Of Citations
Isafe [5]	✓	✗	Videos	Happiness, Sad, Surprise, Disgust, Fear, Angry, Uncertain And No Emotion	395 Clips	44 Volunteers	✓	12
IFED [21]	✓	✗	Videos	Anger, Contempt, Disgust, Fear, Happiness, Sadness And Surprise	N/A	112	✓	N/A
HEU Emotion [22]	✗	✓	Facial Expression, Body Posture And Emotional Speech	10 Emotions	19,004 Video Clips	From Movies, TV Series, And Variety Shows	✗	14
Toronto Emotional Speech Set (TESS) Dataset [23]	✓	✗	Audio	Anger, Disgust, Fear, Happiness, Pleasant Surprise, Sadness, And Neutral	2800 Audio Files	200 Target Words	✗	06
ASVP-ESD(Speech & Non-Speech Emotional Sound) [24]	✓	✗	Audio	Boredom, Neutral, Happiness, Sadness, Anger, Fear, Surprise, Disgust, Excite, Pleasure, Pain And Disappointment	5146 Audio Files	From Movies, Tv Show, YouTube and Others	✗	N/A
SAVEE [25]	✗	✓	Audio And Video	Six Basic Emotions And Neutral	480 British English Utterances	4 Male Actors	✗	171
Emotional Speech Database (ESD) [26]	✓	✗	Audio	Neutral, Happy, Angry, Sad And Surprise	350 Parallel Utterances	10 Native Mandarin Speakers, And 10 English Speakers	✗	30

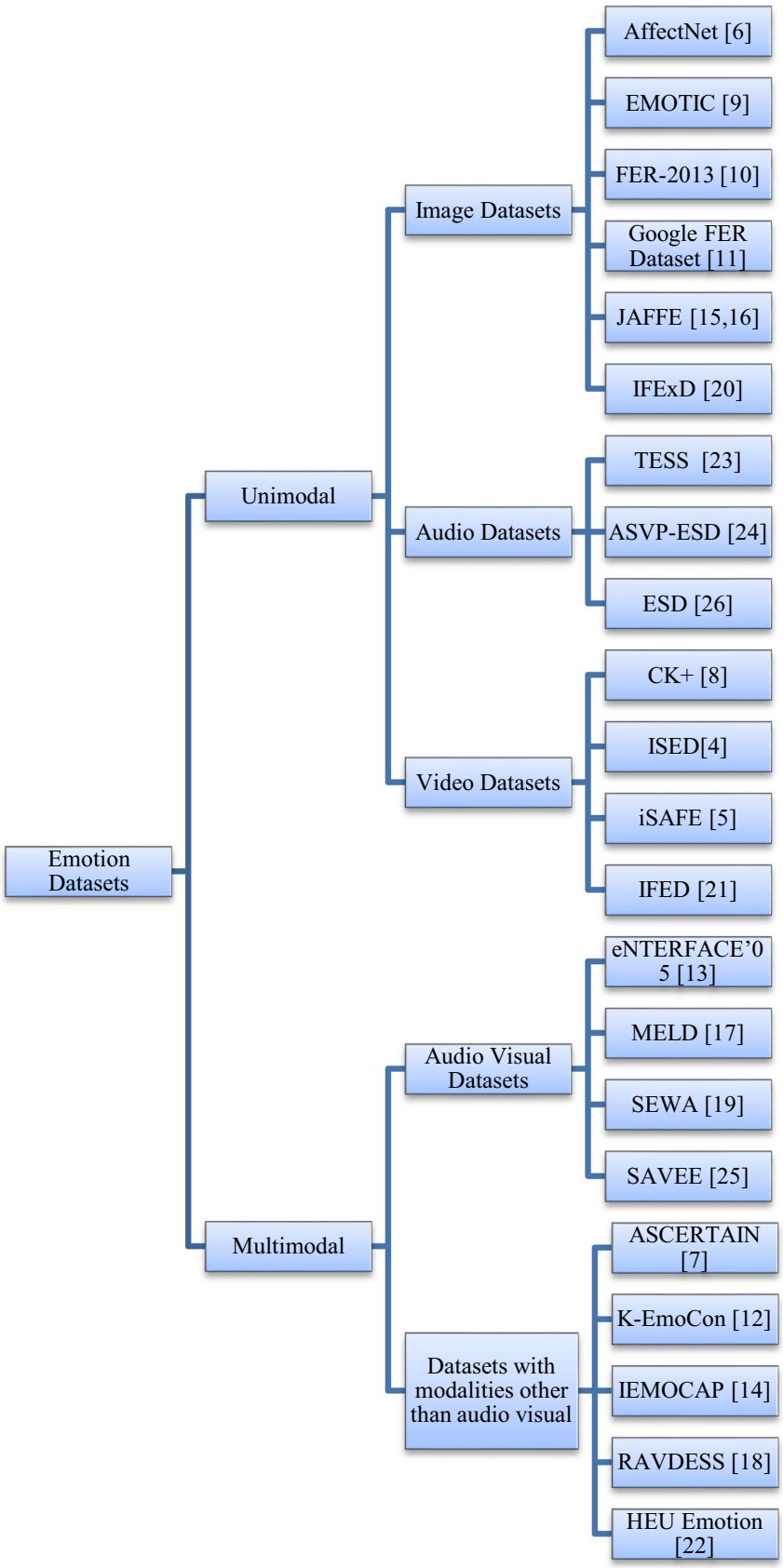

Fig. 1 The proposed taxonomy for the surveyed datasets

improving their benchmark accuracy. Furthermore, the literature on lightweight deep learning models for Facial Emotion Recognition (FER) primarily focuses on models with 10 convolution layers, suggesting a need for more computationally efficient models.

To address these gaps, this paper proposes a novel lightweight deep learning model for Indian Emotion Recognition through facial expressions. The proposed model aims to improve accuracy while reducing training time by offering lesser computational complexity compared to existing lightweight model in the literature. By leveraging the unique characteristics of Indian facial expressions, the model intends to enhance the benchmark accuracy of the available Indian datasets [4, 5].

3 Suggested methodology

The proposed methodology seeks to improve the baseline findings established by the ISED [4] and iSAFE [5] datasets with reduced computational complexity using the proposed lightweight models and also advises evaluating the results. The proposed system's architecture is shown in Fig. 2. To eliminate background and extract faces, the suggested method uses Haar Cascade on the input image. The second stage concentrates on extracting crucial features like position of chin, lips, and eyebrows for accurately distinguishing emotions. In stage 3, the backdrop is blurred once these features have been extracted and marked on the source image. Proposed deep learning models are applied in the final stage of classification. Section 3.1 below describes different models proposed in this paper for classification.

3.1 Proposed classification models

To increase the baseline accuracy with reduced computational time, the paper proposes a custom built CNN model and a custom CNN with SVM model. Three specially designed CNN architectures are presented in this research and are shown in Fig. 3. The proposed architectures are designed with a reduced number of convolution layers to enhance computational efficiency while maintaining high accuracy. Compared to [36], which has 10 convolution layers in their model for emotion recognition, the proposed model is designed with only 4 convolution layers along with two max pooling layers. Figure 3a describes the custom built CNN having four Convolution layer with Relu activation function along with two fully connected dense layers. Max Pooling is employed in the suggested architecture. Figure 3b shows the architecture of custom built CNN where the dense layers were replaced with SVM after flattening. As described in Fig. 2, both of these architectures (Fig. 3a and b) are applied after extracting and highlighting important features of the input image. The dimensions of the original image were reduced from 256×256 to 200×200 after feature extraction. The third architecture applies custom built CNN on the input image directly without any feature extraction as shown in Fig. 3c. The paper also applies Resnet50 model for the purpose of comparison.

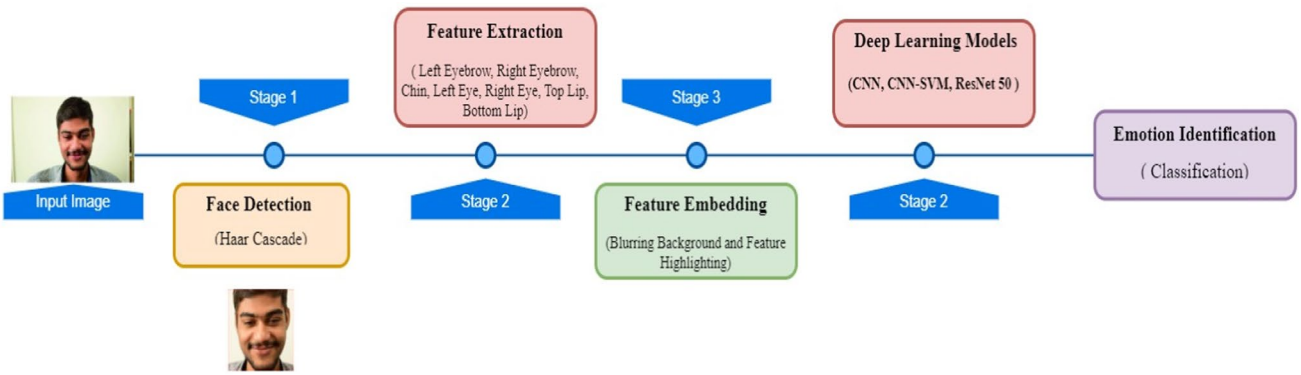

Fig. 2 Architectural Diagram of Proposed System

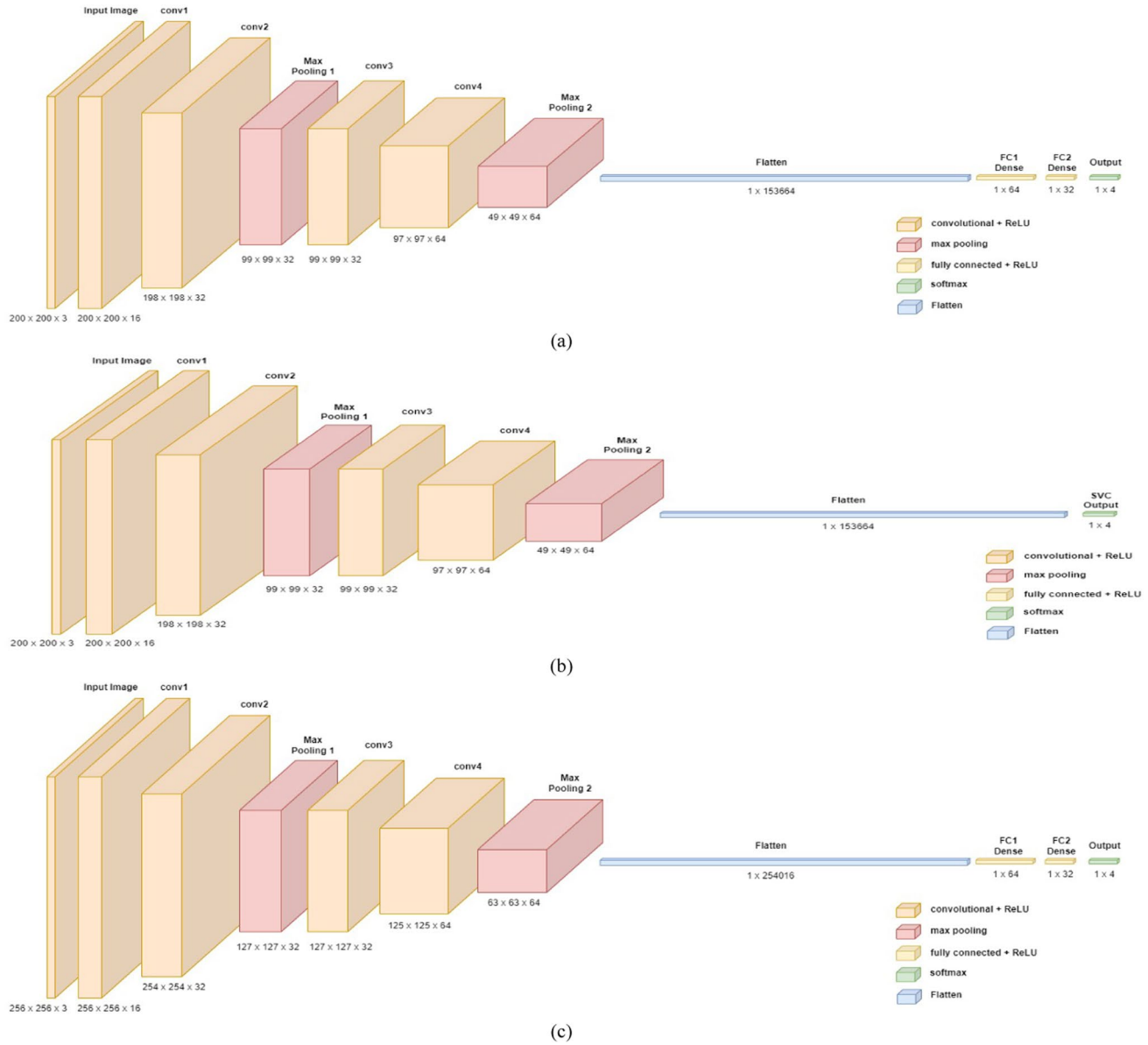

Fig. 3 **a** Architecture of Custom Built CNN with Manual Feature Extraction, **b** Architecture of Custom Built CNN with SVM after Manual Feature Extraction, **c** Architecture of Custom Built CNN without Manual Feature Extraction

4 Experiments and results

The experiments employed the datasets ISED [4] and iSAFE [5] to confirm the efficacy of the models suggested in this research. Accuracy is an evaluation measure used to assess recognition effects.

4.1 Datasets employed

4.1.1 The Indian spontaneous expression database (ISED)

ISED is a single-modality video dataset of Indian-natural face expressions. The dataset contains 428 video clips of 50 healthy volunteers, including 29 men and 21 women, who are between the ages of 18 and 22 and come from various

parts of India. Four emotions—happiness, surprise, sadness, and disgust—were noted in the samples. The authors of this dataset applied Principal Component Analysis (PCA) with Linear Discriminant Analysis (LDA) after extracting features using the Local Gabor Binary Pattern (LGBP), for emotion recognition of Indians and achieved 86.46% accuracy.

4.1.2 Indian semi-acted facial expression (iSAFE)

The dataset consists of 395 clips of 44 volunteers between the ages of 17 and 22; their facial expressions were captured as they watched a few stimulating videos; the volunteers self-annotated the facial movements, which were then cross-annotated by observers for the categories of happiness, sadness, surprise, disgust, fear, anger, uncertainty, and no emotion. The authors used the CNN ResNet34 model to establish a baseline and attained 92% accuracy.

4.2 Experimental results

The datasets are divided into training set, validation set and test set. During data pre-processing, the video is divided into image frames and peak intensity RGB image of 256×256 dimensions is taken as input. Data augmentation was carried out for ISED dataset [4] by varying brightness and contrast. Feature extraction was carried out by identifying landmarks as left eyebrow, right eyebrow, chin, left eye, right eye, upper lip and lower lip. Figure 4 shows facial image after feature extraction and outlining.

After Image extraction and outlining, it is given as an input to different models proposed in the paper. Figure 5a and b describes parameter summary of custom built CNN model with manual feature extraction and without manual feature extraction respectively.

According to the findings of experiments performed on the ISED dataset [4], the custom CNN models with manual feature extraction and those without manual feature extraction respectively offered accuracy of 97.6% and 97.01%. Whereas custom CNN with SVM after feature extraction provided accuracy of 94.87% and Resnet 50 model trained on Imagenet dataset provided accuracy of 97.33%. For each proposed model, a confusion matrix is provided in Fig. 6. To further evaluate the proposed models, different evaluation metrics like precision, recall and F1 score are also applied. These evaluation parameters and their corresponding values for the ISED dataset are described in Table 2. The proposed lightweight CNN models on an average took approximately 80 s of training time whereas the resnet50 model took approximately 400 s to converge on ISED dataset.

The results of iSAFE dataset [5] indicate that custom built CNN model with manual Feature extraction provided accuracy of 96.72% and Custom CNN without manual feature extraction provided accuracy of 92.45% whereas CNN with SVM after manual feature extraction and RESNET 50 provided accuracy of 95.37% and 96.48% respectively. Figure 7 provides confusion matrix for all proposed models. Different assessment criteria, such as precision, recall, and F1 score, are also used to further assess the suggested models. Table 3 provides a description of these evaluation parameters and their associated values for iSAFE dataset. The proposed lightweight CNN models on an average took approximately 860 s of training time whereas the Resnet50 model took approximately 920 s to converge on iSAFE dataset.

Table 4 presents a comparative analysis of accuracies achieved by the proposed lightweight custom-built CNN models, along with their corresponding training times, for the ISED and iSAFE datasets. The results highlight significant

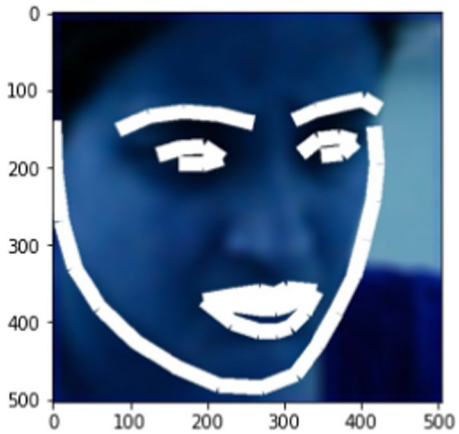

Fig. 4 Feature Extraction and outlining

Layer (type)	Output Shape	Param #
conv2d (Conv2D)	(None, 200, 200, 16)	448
conv2d_1 (Conv2D)	(None, 198, 198, 32)	4640
max_pooling2d (MaxPooling2D)	(None, 99, 99, 32)	0
conv2d_2 (Conv2D)	(None, 99, 99, 32)	9248
conv2d_3 (Conv2D)	(None, 97, 97, 64)	18496
max_pooling2d_1 (MaxPooling2D)	(None, 49, 49, 64)	0
flatten (Flatten)	(None, 153664)	0
dense (Dense)	(None, 64)	9834560
dense_1 (Dense)	(None, 32)	2080
dense_2 (Dense)	(None, 4)	132

Total params: 9,869,604
Trainable params: 9,869,604
Non-trainable params: 0

(a)

Layer (type)	Output Shape	Param #
conv2d_8 (Conv2D)	(None, 256, 256, 16)	448
conv2d_9 (Conv2D)	(None, 254, 254, 32)	4640
max_pooling2d_4 (MaxPooling2D)	(None, 127, 127, 32)	0
conv2d_10 (Conv2D)	(None, 127, 127, 32)	9248
conv2d_11 (Conv2D)	(None, 125, 125, 64)	18496
max_pooling2d_5 (MaxPooling2D)	(None, 63, 63, 64)	0
flatten_2 (Flatten)	(None, 254016)	0
dense_6 (Dense)	(None, 64)	16257088
dense_7 (Dense)	(None, 32)	2080
dense_8 (Dense)	(None, 4)	132

Total params: 16,292,132
Trainable params: 16,292,132
Non-trainable params: 0

(b)

Fig. 5 Parameter Summary for Custom built CNN **a** with Manual Feature Extraction, **b** Without Manual Feature extraction

improvements in accuracy compared to the baseline. The lightweight custom-built CNN model with manual feature extraction demonstrates the most remarkable accuracy improvement of 11.14% for the ISED dataset and 4.72% for the iSAFE dataset. The CNN model without manual feature extraction also exhibits good accuracy improvement of 10.55% for the ISED dataset but provides a relatively smaller improvement of 0.45% for the iSAFE dataset. On the other hand, the CNN model with SVM after manual feature extraction shows moderate accuracy improvement, with 8.41% for the ISED dataset and 3.37% for the iSAFE dataset. The ResNet-50 model achieves accuracy improvements of 10.87% for the ISED dataset and 4.48% for the iSAFE dataset. Notably, the lightweight CNN model with manual feature extraction outperforms the ResNet-50 model in terms of accuracy improvement. Furthermore, the training time for the lightweight CNN model is significantly lower, with a difference of 320 s for the ISED dataset and 60 s for the iSAFE dataset compared to the ResNet-50 model. Figure 8a and b illustrate the accuracy comparison graphs of each suggested model with the baseline for the ISED and iSAFE datasets, respectively.

5 Conclusion

This paper proposed an approach for precise human emotion recognition among Indians using custom-built lightweight CNN model. By utilizing the ISED and iSAFE datasets, which contain video recordings of Indians in the age group of 17–22 years, we were able to extract key facial features and apply these models for accurate emotion recognition. The experimental results demonstrate significant improvements in accuracy, with the lightweight CNN model with manual feature extraction achieving the most remarkable improvement of 11.14% for ISED dataset and 4.72% for iSAFE dataset, outperforming the baseline models. Comparative analysis with the widely used ResNet-50 model reveals that our proposed CNN model with manual feature extraction provides competitive accuracy improvements of 0.27% for the ISED dataset and 0.24% for the iSAFE dataset, while requiring significantly less training time. This highlights the efficiency and effectiveness of the custom-built lightweight CNN model in capturing the unique emotional expressions of individuals in the Indian population. Furthermore, the inclusion of additional evaluation metrics such as precision, recall, and F1 score enhances the analysis of model performance. Although a direct comparison of these metrics with the baseline is not possible due to the lack of availability of these in the literature, the experimental findings strongly support the superiority of our proposed models in accurately recognizing human emotions. As a future direction, we plan to expand the evaluation of our proposed models by testing them on diverse

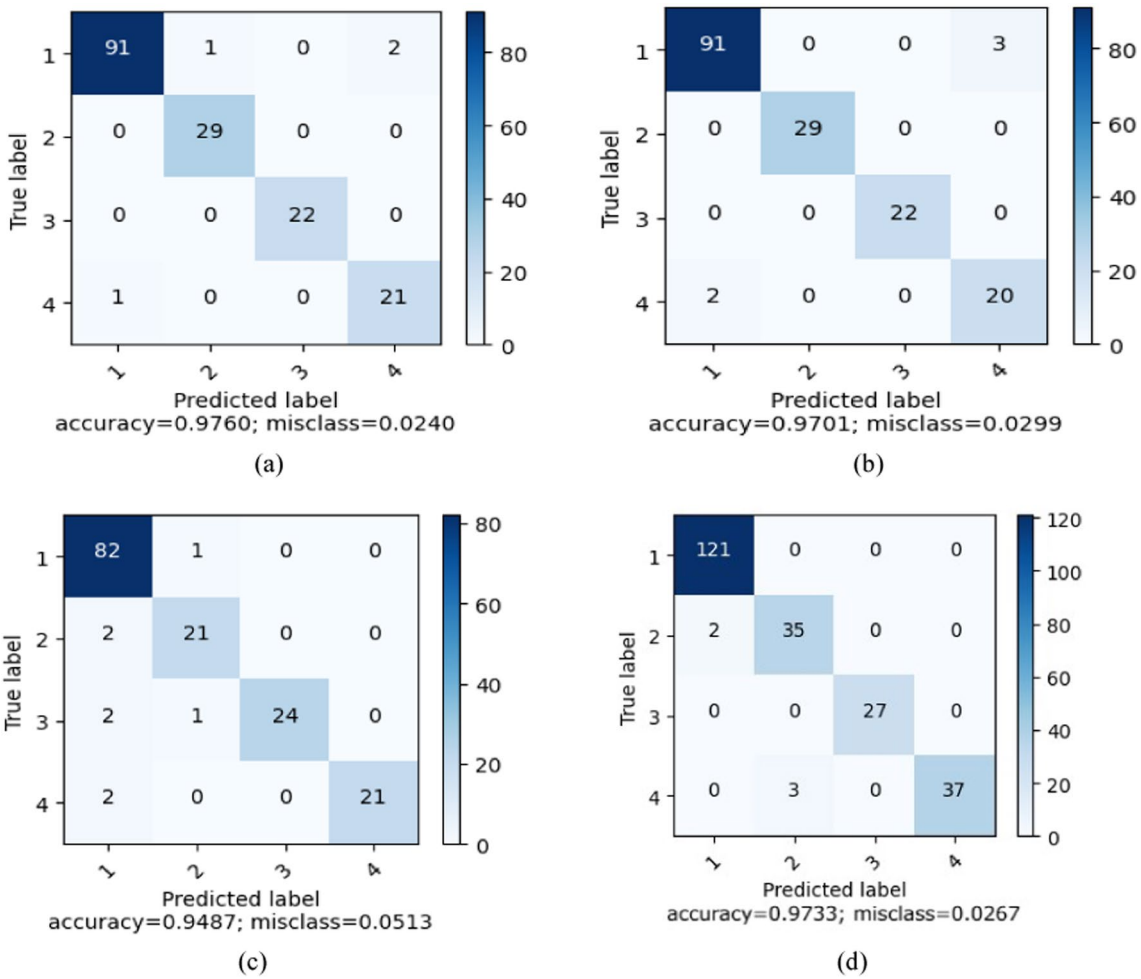

Fig. 6 Confusion Metrics for **a** Custom Built CNN with manual feature extraction **b** Custom Built CNN without manual feature extraction **c** Custom Built CNN with SVM after manual feature extraction **d** Resnet 50 on ISED dataset

Table 2 Evaluation metrics for ISED dataset

Proposed models	Metrics	Happiness	Surprise	Sadness	Disgust
CNN with manual features	F1-Score	0.9785	0.9831	1.0000	0.9333
	Precision	0.9891	0.9667	1.0000	0.9130
	Recall	0.9681	1.0000	1.0000	0.9545
CNN without manual features	F1-Score	0.9733	1.0000	1.0000	0.8889
	Precision	0.9785	1.0000	1.0000	0.8696
	Recall	0.9681	1.0000	1.0000	0.9091
CNN-SVM with manual features	F1-Score	0.9590	0.9130	0.9411	0.9545
	Precision	0.9318	0.9130	1.0000	1.0000
	Recall	0.9880	0.9130	0.8889	0.9130
Resnet 50	F1-Score	99.18	93.33	100	96.1
	Precision	98.37	92.11	100	100
	Recall	100	94.59	100	92.5

image datasets that encompass various ethnicities, with a particular focus on Asian datasets. This will enable us to assess the generalizability of our approach and further validate its effectiveness in different cultural contexts.

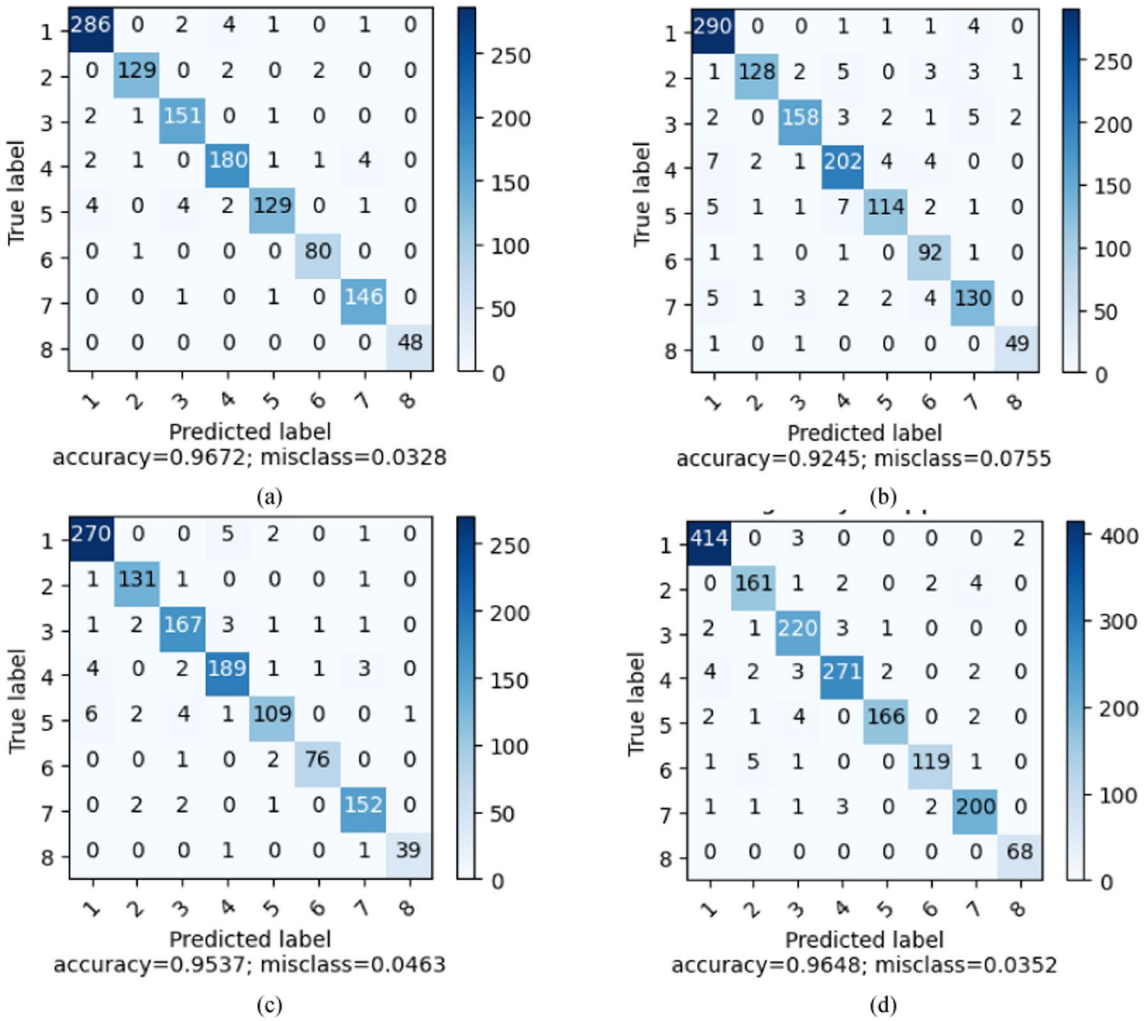

Fig. 7 Confusion Metrics for **a** Custom Built CNN with manual feature extraction **b** Custom Built CNN without manual feature extraction, **c** Custom Built CNN with SVM after manual feature extraction and **d** Resnet 50 Model on iSAFE Dataset

Table 3 Evaluation metrics for iSAFE dataset

Proposed Models	Metrics	Happiness	Sadness	Surprise	Disgust	Fear	Anger	Uncertain	No emotion
CNN with manual features	F1-Score	0.97279	0.97358	0.96486	0.95491	0.94505	0.97561	0.97333	1.00000
	Precision	0.97279	0.97727	0.95570	0.95745	0.96992	0.96386	0.96053	1.00000
	Recall	0.97279	0.96992	0.97419	0.95238	0.92143	0.98765	0.986486	1.00000
CNN without manual features	F1-Score	0.95238	0.92754	0.93215	0.91610	0.89764	0.90640	0.89347	0.95146
	Precision	0.92949	0.96241	0.95181	0.91403	0.92683	0.85981	0.90278	0.94231
	Recall	0.97643	0.89510	0.91329	0.91818	0.87023	0.95833	0.88435	0.96078
CNN with SVM after manual features	F1-Score	0.96430	0.96680	0.94620	0.94740	0.91210	0.96820	0.96200	0.96300
	Precision	0.95740	0.95620	0.94350	0.94970	0.93970	0.97440	0.95600	0.97500
	Recall	0.97120	0.97760	0.94890	0.94500	0.88620	0.96200	0.96820	0.95120
Resnet 50	F1-Score	0.9822	0.9443	0.9565	0.9627	0.9651	0.952	0.9592	0.9855
	Precision	0.9764	0.9415	0.9442	0.9713	0.9822	0.9675	0.9569	0.9714
	Recall	0.9881	0.9471	0.9692	0.9542	0.9486	0.937	0.9615	1

Table 4 Comparative analysis of proposed lightweight CNN models

Datasets employed	Models	Accuracy %	Enhanced accuracy compared to baseline %	Training time	Improvement in training time requirement
ISED	CNN with manual features	97.60	11.14	Approx. 80	320
	CNN without manual features	97.01	10.55		
	CNN with SVM after manual features	94.87	8.41		
	RESNET 50	97.33	10.87	400	
iSAFE	CNN with manual features	96.72	4.72	Approx. 860	60
	CNN without manual features	92.45	0.45		
	CNN with SVM after manual features	95.37	3.37		
	RESNET 50	96.48	4.48	920	

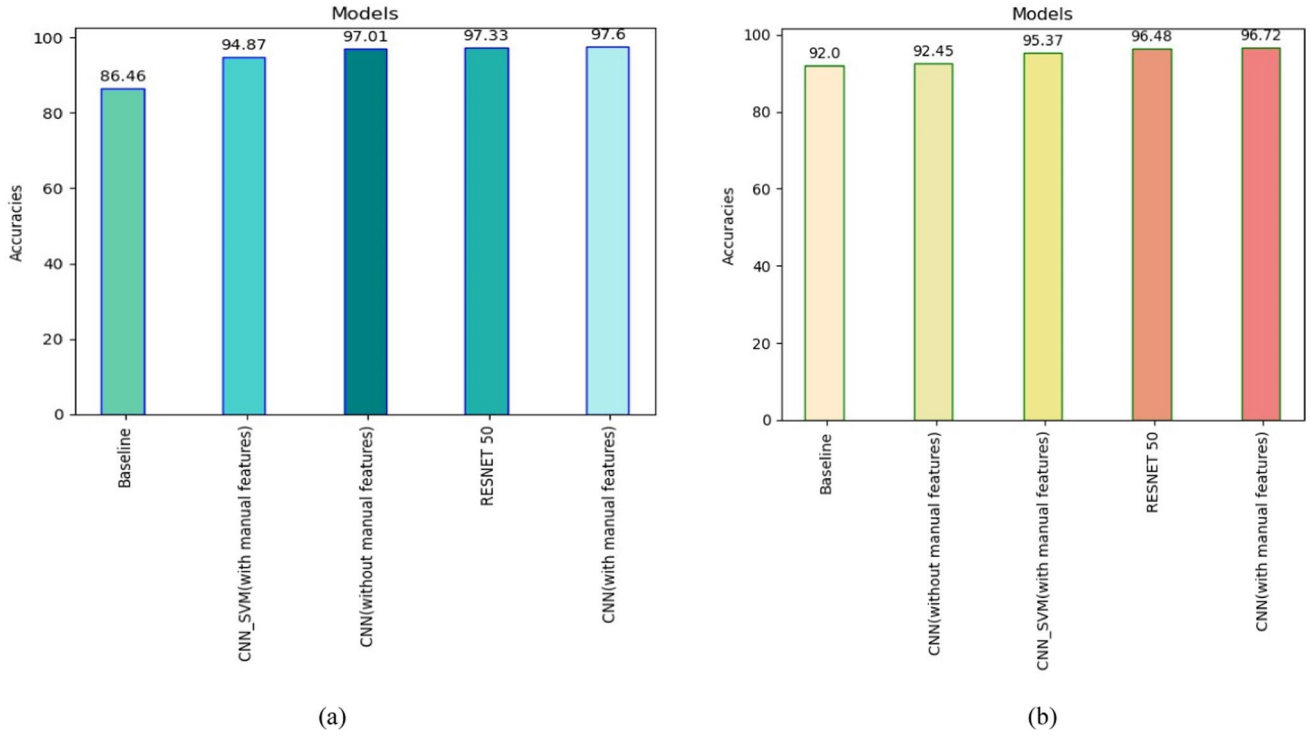

Fig. 8 Accuracy Comparison of proposed models with baseline results for **a** ISED dataset **b** iSafe Dataset

Author contributions Ruhina Karani: Study conception and Design, Analysis and interpretation of results, draft manuscript preparation Jay Jani: Analysis and interpretation of results, draft manuscript preparation Dr. Sharmishta Desai: Analysis and interpretation of results, draft manuscript review.

Data availability The dataset used in this study cannot be made publicly available due to restrictions on data sharing imposed by the data owner.

Declarations

Competing interests The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Yongjun X, Liu X, Cao X, Huang C, Liu E, Qian S, Liu X, Yanjun W, Dong F, Qiu C-W, Qiu J, Hua K, Wentao S, Jian W, Huiyu X, Han Y, Chenguang F, Yin Z, Liu M, Roepman R, Dietmann S, Virta M, Kengara F, Zhang Z, Zhang L, Zhao T, Dai J, Yang J, Lan L, Luo M, Liu Zhaofeng, An T, Zhang B, He X, Cong S, Liu X, Zhang W, Lewis JP, Tiedje JM, Wang Q, An Z, Wang F, Zhang L, Huang T, Chuan L, Cai Z, Wang F, Zhang J. Artificial intelligence: a powerful paradigm for scientific research". Innovation. 2021;2(4):100179. https://doi.org/10.1016/j.xinn.2021.100179.
2. Karani R, Desai S. Review on multimodal fusion techniques for human emotion recognition. Int J Adv Comput Sci Appl (IJACSA). 2022. https://doi.org/10.14569/IJACSA.2022.0131035.
3. Chaudhari A, Bhatt C, Nguyen TT, et al. Emotion recognition system via facial expressions and speech using machine learning and deep learning techniques. SN Comput Sci. 2023;4:363. https://doi.org/10.1007/s42979-022-01633-9.
4. Happy SL, Patnaik P, Routray A, Guha R. The Indian spontaneous expression database for emotion recognition. IEEE Trans Affect Comput. 2015;8:1–1. https://doi.org/10.1109/TAFFC.2015.2498174.
5. Singh S, Benedict S. Indian semi-acted facial expression (iSAFE) dataset for human emotions recognition. Singapore: Springer Singapore; 2019.
6. Mollahosseini A, Hasani B, Mahoor M. AffectNet: a database for facial expression, valence, and arousal computing in the wild. IEEE Trans Affect Comput. 2017. https://doi.org/10.1109/TAFFC.2017.2740923.
7. Subramanian R, Wache J, Abadi MK, Vieriu RL, Winkler S, Sebe N. ASCERTAIN: emotion and personality recognition using commercial sensors. IEEE Trans Affect Comput. 2018;9(2):147–60. https://doi.org/10.1109/TAFFC.2016.2625250.
8. P Lucey, JF Cohn, T Kanade, J Saragih, Z Ambadar, I Matthews. "The extended Cohn-Kanade dataset (CK+): a complete dataset for action unit and emotion-specified expression." 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition—Workshops, San Francisco, CA, USA. 2010. 94–101
9. R Kosti, JM Alvarez, A Recasens, A Lapedriza. "EMOTIC: emotions in context dataset," 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Honolulu, HI, USA, 2017;2309–17. doi: https://doi.org/10.1109/CVPRW.2017.285.
10. Goodfellow IJ, Erhan D, Carrier PL, Courville A, Mirza M, Hamner B, et al. Challenges in representation learning: a report on three machine learning contests. Neural Netw. 2015;64:59–63. (**Special issue on deep learning of representations**).
11. Vemulapalli R, Agarwala A. A compact embedding for facial expression similarity. In: 2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR), 2019. p. 5676–85.
12. Park CY, Cha N, Kang S, et al. K-EmoCon, a multimodal sensor dataset for continuous emotion recognition in naturalistic conversations. Sci Data. 2020;7:293. https://doi.org/10.1038/s41597-020-00630-y.
13. O Martin, I Kotsia, B Macq, I Pitas. "The eNTERFACE' 05 audio-visual emotion database," 22nd International Conference on Data Engineering Workshops (ICDEW'06), Atlanta, GA, USA, 2006;8–8. https://doi.org/10.1109/ICDEW.2006.145.
14. Busso C, Bulut M, Lee CC, et al. IEMOCAP: interactive emotional dyadic motion capture database. Lang Resour Evaluat. 2008;42:335–59. https://doi.org/10.1007/s10579-008-9076-6.
15. Lyons MJ, Kamachi M, Gyoba J. Coding facial expressions with gabor wavelets (IVC Special Issue). arXiv. 2020. https://doi.org/10.48550/arXiv.2009.05938.
16. Lyons MJ. "Excavating AI" re-excavated: debunking a fallacious account of the JAFFE dataset. arXiv. 2021. https://doi.org/10.48550/arXiv.2107.13998.
17. Poria S, Hazarika D, Majumder N, Naik G, Cambria E, Mihalcea R. MELD: a multimodal multi-party dataset for emotion recognition in conversations. In proceedings of the 57th annual meeting of the association for computational linguistics. Florence: Association for Computational Linguistics; 2019. p. 527–36.
18. Livingstone SR, Russo FA. The Ryerson audio-visual database of emotional speech and song (Ravdess): a dynamic, multimodal set of facial and vocal expressions in north American English. PLoS ONE. 2018;13(5):e0196391. https://doi.org/10.1371/journal.pone.0196391.
19. Kossaifi J, Walecki R, Panagakis Y, Shen J, Schmitt M, et al. SEWA DB: a rich database for audio-visual emotion and sentiment research in the wild. IEEE Trans Pattern Anal Mach Intell. 2021;43(3):1022–40.
20. IFEXD. (n.d.). IFEXD. Retrieved from https://ifexd.github.io/index.
21. Arunashri, Venkateshu K V, Lalitha C. A database for facial expressions among Indians. MedPulse Int J Anat. 2021; 17(2): 12–16. http://www.medpulse.in/Anatomy.
22. Jing C, Chenhui W, Kejun W, Chaoqun Y, Zhao Cong X, Tao ZX, Ziqiang H, Meichen L, Yang T. HEU Emotion: a large-scale database for multimodal emotion recognition in the wild. Neural Comput Appl. 2021. https://doi.org/10.1007/s00521-020-05616-w.
23. Pichora Fuller MK, Dupuis K. Toronto emotional speech set (TESS) (DRAFT VERSION). Scholars Portal Dataverse. 2020. https://doi.org/10.5683/SP2/E8H2MF.
24. Landry DTT, He Q, Yan H, Li Y. ASVP-ESD: a dataset and its benchmark for emotion recognition using both speech and non-speech utterances. Global Sci J. 2020;8(5):1793.
25. Jackson P, Ul Haq S. (2011). Surrey Audio-Visual Expressed Emotion (SAVEE) database. Guildford (UK): University of Surrey; 2014.
26. Zhou K, Sisman B, Liu R, Li H. Emotional voice conversion: Theory, databases and ESD. Speech Commun. 2022;137:1–18.

27. Schoneveld L, Othmani A, Abdelkawy H. Leveraging recent advances in deep learning for audio-Visual emotion recognition. Pattern Recognit Lett. 2021;146:1–7. https://doi.org/10.1016/j.patrec.2021.03.007.
28. Farhoudi Z, Setayeshi S. Fusion of deep learning features with mixture of brain emotional learning for audio-visual emotion recognition. Speech Commun. 2021;127:92–103. https://doi.org/10.1016/j.specom.2020.12.001.
29. Babajee P, Suddul G, Armoogum S, Foogooa R. Identifying human emotions from facial expressions with deep learning. Zoom Innovat Consumer Technol Conf (ZINC). 2020. https://doi.org/10.1109/ZINC50678.2020.9161445.
30. Lee S, Han DK, Ko H. Multimodal emotion recognition fusion analysis adapting BERT with heterogeneous feature unification. IEEE Access. 2021;9:94557–72. https://doi.org/10.1109/ACCESS.2021.3092735.
31. Lee S, Han DK, Ko H. Multimodal emotion recognition fusion analysis adapting BERT with heterogeneous feature unification. IEEE Access. 2021;9:94557–72. https://doi.org/10.1109/ACCESS.2021.3092735.
32. Darapaneni R, Choubey P, Salvi A, Pathak SS, Paduri AR. Facial expression recognition and recommendations using deep neural network with transfer learning. 11th IEEE Ann Ubiquitous Comput Electron Mobile Commun Conf (UEMCON). 2020. https://doi.org/10.1109/UEMCON51285.2020.9298082.
33. Supta SR, Sahriar MR, Rashed MG, Das D, Yasmin R. An effective facial expression recognition system. IEEE Int Women Eng (WIE) Conf Electrical Comput Eng (WIECON-ECE). 2020. https://doi.org/10.1109/WIECON-ECE52138.2020.9397965.
34. Zhang X, Wang M-J, Guo X-D. Multi-modal emotion recognition based on deep learning in speech, video and text. IEEE 5th Int Conf Signal Image Proc (ICSIP). 2020. https://doi.org/10.1109/ICSIP49896.2020.9339464.
35. Nemati S. Canonical correlation analysis for data fusion in multimodal emotion recognition. Int Symposium Telecommun (IST). 2018. https://doi.org/10.1109/ISTEL.2018.8661140.
36. Barros P, Churamani N, Sciutti A. The FaceChannel: a fast and furious deep neural network for facial expression recognition. SN Comput Sci. 2020;1:321. https://doi.org/10.1007/s42979-020-00325-6.