Mobilenetv3: a deep learning technique for human face expressions identification

Article *in* International Journal of Information Technology · August 2023

DOI: 10.1007/s41870-023-01380-x

2 authors, including:

Babu rajendra prasad Singothu

4 PUBLICATIONS **34** CITATIONS

SEE PROFILE

Int. j. inf. tecnol.
https://doi.org/10.1007/s41870-023-01380-x

ORIGINAL RESEARCH

Mobilenetv3: a deep learning technique for human face expressions identification

S. Babu Rajendra Prasad[1] · **B. Sai Chandana**[1]

Received: 24 January 2023 / Accepted: 17 July 2023

Abstract One of the most cutting-edge methods for understanding a person's current psychological state is emotion recognition. Due to the nature of emotional expression, an effective feature extractor and classifier are required. Thermal imagery is the way of taking images in the infrared spectrum that depend on thermal emissivity. But when taken in places with strong sunlight and fog, thermal images may not produce clear results. To overcome all of the issues in human emotion recognition, we proposed a deep learning techniques. In this paper, we propose a novel deep-learning technique of MobileNetv3 to classify the face emotions in thermal images. It has five steps to recognize and classify the emotions on human faces. Initially, the given input image is using the mean filter to remove the noise and then normalize the data using the min–max normalization method. After that, the Binary Dragonfly Algorithm (BDA) is utilized to extract the features such as forehead, eye, chick (left and right), nose, and mouth on a human face. Then, segment the human areas using Enhanced Crow Search Algorithm (ECSA). Then, utilizing the Dense-UNet technique to recognize the face in given thermal face images. Finally, classify the face into expressions such as neutral, sad, fearful, happy, contempt, surprise, laughing, and anger utilizing the MobileNetv3 technique. The performance of experiments using two different thermal face expression databases, the IR database and the IRIS Thermal/Visible database for good performance with precision, accuracy, F1-score, and recall parameters. The proposed approach outperforms "state-of-the-art" techniques for face expression detection on thermal images and categorizes the items.

Keywords Thermal imagery · Human emotion recognition · MobileNetv3 · Noise removal · Classification · Feature extraction

1 Introduction

Automatic facial emotion detection has gained increased interest in recent years as a result of the many applications it can be used. The majority of work on facial expression detection uses visible spectrum images, but these images are susceptible to changes in lighting, which can affect how well the emotion detection techniques operate [1–3]. Because thermal imaging measures facial temperatures, it is unaffected by changes in lighting. This feature suggests that in some applications and circumstances, such as nighttime or low-light daytime recordings, emotion recognition using thermal face images may be more practical. In affective computing, expression recognition is a critical issue for a deeper knowledge of human behavior. It enables more effective and better communication between people and computers [4, 5]. The majority of earlier works used visual data from the face image to complete the task of emotion recognition. The most prevalent source of visual data, the visible light spectrum can produce variations in the facial images depending on the illumination. However, the huge differences in texture and appearance between images could result in extremely low identification performance [6–9]. Additionally, as security and privacy concerns are receiving more attention in today's society, visual data of human faces are more susceptible to this problem. However, infrared thermal images, which are

✉ B. Sai Chandana
saichandanas869@gmail.com

S. Babu Rajendra Prasad
baburajendraprasad655@gmail.com

[1] School of CSE, VIT-AP University, Amaravathi, Vijayawada, Andhra Pradesh, India

Published online: 03 August 2023

unaffected by the problems identified can work as a substitute source for identifying facial expressions.

By measuring the heat irradiation of the body, thermal imaging enables non-invasive and non-contact surveillance of the skin temperature [10–12]. These temperature fluctuations in the human body are brought on by emotional perspiration and blood flow within veins. For example, worry, being startled, or feeling embarrassed have all been studied using thermal imaging. The comparison of concurrently acquired data by thermal imaging and by gold standard methods demonstrated the validity and reliability of thermal imaging as the platform for the study of emotional states [13, 14]. Even while the work of recognizing facial emotions in the thermal spectrum has become more and more popular recently, it is still difficult. The lack of available data is one of the primary causes.

There are currently very few datasets of thermal spectrum facial images that have emotion classifications. In contrast, some significant datasets of visible images for emotion detection have become accessible in recent years, including EmotiW, EmotioNet, and FER2013 [15]. Here, we proposed a deep learning strategy to resolve all the problems in thermal images for recognizing human facial expressions. In this research, we proposed MobileNetv3 a unique deep-learning technique. First, the noise in the provided input image is removed using a mean filter, and then the data is normalized using the min–max normalization method. The BDA approach is then used to extract the features of the human face, including the forehead, eye, chick (left and right), nose, and mouth. Then, using the Enhanced Crow Search Algorithm (ECSA) to segment the human areas. Using the Dense-UNet method to identify the face in thermal face images. Finally, employing the MobileNetv3 technique to categorize the face into its expressions, such as neutral, sad, fearful, happy, contemptuous, surprised, laughing, and angry. The execution of studies using the IR database and IRIS Thermal/Visible database, two distinct thermal face expression databases. The key contribution is,

- To remove the noise from thermal face images, utilized a mean filter, and also normalize the data using the min–max normalization method.
- Utilizing the Binary Dragonfly Algorithm (BDA) technique to extract the features from the human face such as the forehead, eye, chick (left and right), nose, and mouth.
- An enhanced crow search algorithm is utilized to segment the human as binary segmentation in given thermal human images. Then using the Dense-UNet method to detect the human face for better classification.
- To classify the expressions on a human face, we utilize the MobileNetv3 technique. It classifies the expressions into neutral, sad, fearful, happy, contemptuous, surprised, laughing, and angry.
- To evaluate performance, we utilized two different thermal face databases IR database and IRIS Thermal/Visible database for good performance.

The remaining sections of the article are organized as follows. The paper's relevant literature is shown in Sect. 2. In Sect. 3, the proposed method is described. The outcome is presented in Sect. 4. Finally, the conclusions are stated in Sect. 5.

2 Literature survey

Recognizing and categorizing human facial expressions involves a wide range of techniques and methodologies. So, we investigated a few articles on thermal imaging and facial expression identification. To compute the multivariate time-series thermal video sequences and identify human emotion and provide distraction ideas, Nayak et al. [16] suggested a three-stage HCI architecture. A faster R-CNN architecture was employed in the first stage to detect faces, eyes, and noses. The Multiple Instance Learning (MIL) algorithms was used to follow the face ROIs throughout the thermal video. The MTS information was developed by calculating the mean intensity of the region of interest. The DTW approach were utilized to classify the human emotions given by video stimulus in the second step utilizing the smoothed MTS information. In the third step of HCI, the presented work offers pertinent recommendations from a psychological and physical distraction perspective.

IRFacExNet (InfraRed Facial Expression Network) is a deep learning network that has been developed by Bhattacharyya et al. [17] for the recognition of expressions on the face from thermal pictures. It makes use of two building blocks, the Transformation unit, and the Residual unit, to retrieve key information relevant to the facial expressions from the input pictures. The traits that were retrieved make it possible to precisely detect the participants under the study's expressions. To enhance overall performance, the snapshot ensemble method were utilized with a cosine annealing learning rate scheduler. The effectiveness of the suggested model has been assessed using data from the IR database, a publicly accessible dataset created by RWTH Aachen University. The dataset includes the following facial expressions: happy, surprised, sad, neutral, anger, fear, and contempt.

Chopade and Prabhakar [18] suggested a technique for human emotion detection based on wavelet transform and image block patterns (HER-BP-DWT). First of all, notice how the test image has been splitted into several block patterns both vertically and horizontally. Then, sub-blocks are created for each block. The discrete wavelet transform was used to divide the specific region block

into various frequency subbands. Each block's sub bands energies are calculated. It compares the energy of the sub-bands in the reference and test images. This method's major objective is to identify face expressions utilizing a straightforward parameter, like the energy of subbands, which produced via a DWT and simple to apply. Different expressions, such as anger, sadness, happiness, etc., can be efficiently and accurately detected using the suggested HER-BP-DWT approach.

To identify human emotions, Said and Barr [19] presented a face-sensitive CNN (FS-CNN). To identify faces in large-scale images, the suggested FS-CNN was applied. Next, face landmarks are analyzed to forecast expressions for expression recognition. Convolutional neural networks and patch cropping are the two phases that make up the FS-CNN. In the first step, faces in high-resolution pictures are found and cropped for further processing. In the second step, a CNN was utilized to process scale invariance on pyramid images and forecast face expressions based on landmark analytics.

Umer et al. [20] suggested a facial expression recognition system. From the input picture that was gathered, an ROI for face identification was made in the first component. A DL-based CNN design has been created for the second component to retrieve feature learning tasks for categorization purposes to detect the sorts of facial expressions to extract more distinct and discriminating features. In the third component, several data augmentation techniques have been applied to the face picture to enrich the learning parameters of the suggested CNN approach to improve the performance of the suggested system. The trained CNN approach has been fine-tuned in the fourth component through a trade-off between data augmentation and DL features.

3 Problem statement

From above previous paper's limitations are described below,

- Large-sized databases would be needed to ensure the robustness of technologies.
- The complexity of space and the cost of time are high.
- With the face database raise, the face expressions identification rate will suffer. The face databases contain more public databases. Thus, the significant direction of future research is to uphold detection rate stability under high databases.

These all issues motivate us to propose a new methodology to recognize human facial expressions in thermal images.

4 Proposed methodology

Due to its consistency under different lighting conditions, thermal infrared imagery is now being used by several scientists to estimate human emotion. Although infrared imagery is preferred over visible imagery due to its consistent nature for changes in illumination, it is limited in handling eyeglasses because the region around them is dark. In this work, we propose a new MobileNetv3 deep learning method. Recognizing and categorizing the emotions on a person's face involves five processes. The given input image is first processed with a mean filter to remove the noise, and the data is then normalized using the min–max normalization method. The Binary Dragonfly Algorithm (BDA) is then used to extract the features of the human face, including the forehead, eye, cheek (left and right), nose, and mouth. Thereafter utilize the Enhanced Crow Search Algorithm (ECSA) to segment the human areas. Using the Dense-UNet method to detect the face in thermal face images.

Finally, the MobileNetv3 technique will be used to categorize the face into its various expressions, including neutral, sad, fearful, happy, contemptuous, surprised, laughing, and angry. Two separate thermal facial expression databases the IR database and the IRIS Thermal/Visible database were used in the experiments to achieve good performance in terms of precision, accuracy, F1-score, and recall metrics. The proposed methodology's architecture is shown in Fig. 1.

4.1 Pre-processing

Remove the noise from the images and normalize images during the pre-processing stage. To improve the recognition of human actions, we first normalize the image and then eliminate noise from the input images. Here, the noise from the input images is first reduced using mean filtering. A typical technique for image de-noising in the spatial domain involves using several image smoothing templates for image convolution processing to decrease or eliminate noise. The underlying concept of mean filtering is to replace a pixel's solitary grey value with the sum of the grey values of all of its surrounding pixels. After smoothing and mean filtering, the image is known as g(x, y), and it is calculated for a pixel point (x, y) in a given image with f (x, y), where its neighborhood S consists of M pixels, using the method below:

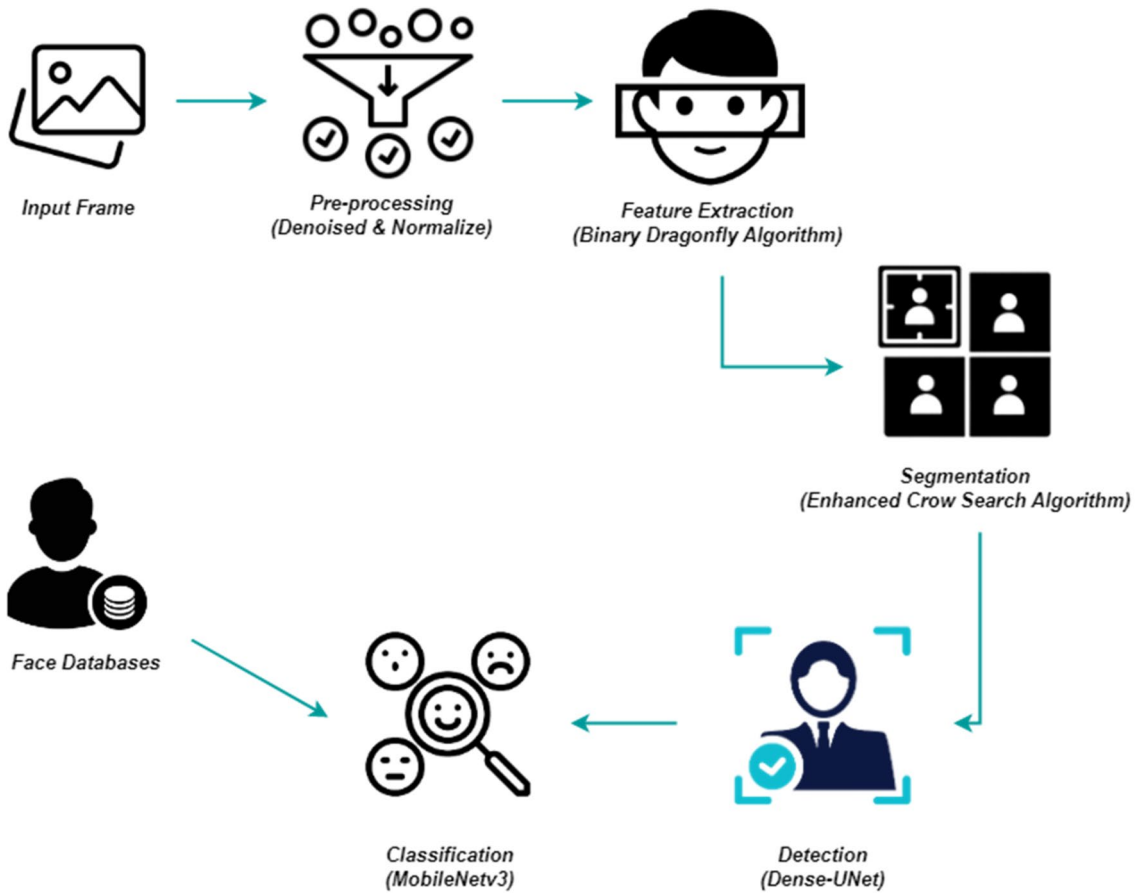

Fig. 1 The proposed methodology architecture

$$G(x,y) = \frac{1}{M} \sum_{(i,j \in S)} f(x,y)(x,y) \notin S \tag{1}$$

The min–max normalization method is selected to standardize the original data, accelerate model convergence, and boost model accuracy, and is denoted as

$$x = \frac{x - \min(x)}{\max(x) - \min(x)} \tag{2}$$

Thermal images are aligned and encapsulated to a detailed anatomic template as part of the normalizing process. Normalization is required because diverse human expressions necessitate different results, which makes it easier to compare one expression to another and translate the outcomes into a standard feature. Normalization frequently entails using a template and a source image to map discrete subject-space data to a reference space.

4.2 Feature extraction

The input image is proceeds to the feature extraction step after pre-processing. Here, the Binary Dragonfly Algorithm (BDA) method is used to extract features from provided thermal face images, including the forehead, eye, chick (left and right), nose, and mouth. As was already indicated, this programmer imitates the natural swarming behaviors of dragonflies. The communication of dragonflies in avoiding the opponent (the worst solution) and locating the food source serves as a model for the exploitative and exploratory mechanisms of DA (the best solution) [21]. The five primary behaviors used in DA for position updates are alignment, separation, attraction, distraction, and cohesion.

These actions are each explained as follows:

Separation tries to prevent the static collision of the present dragonfly with the nearby dragonfly. Separation can be mathematically stated as follows:

$$S_i = -\sum_{j=1}^{M} X - X_j \tag{3}$$

where X is a dragonfly's location in a D-dimensional space (D is the number of choice variables), X_j denotes the neighboring creature's location, and M denotes the number of nearby creatures.

Velocity matching between individuals in a sub-swarm or swarm is made possible through alignment. The calculation for alignment is as follows:

$$A_i = \frac{\sum_{j=1}^{M} V_j}{M} \tag{4}$$

where M is the entire amount of nearby individuals and V_j is their collective velocity.

The term "cohesion" describes the present individual's movement toward the middle of the group of nearby neighbors. The following definition of cohesion:

$$C_i = \frac{\sum_{j=1}^{M} X_j}{M} - X \tag{5}$$

where M is the total amount of dragonflies in the area and X_j is the location of the neighboring insect.

Individuals in natural swarms separate, align, and cooperate as well as draw toward food sources and divert predators' attention. According to attraction, a person should be drawn to food sources. The attraction is defined mathematically as:

$$F_i = Xf - X \tag{6}$$

where Xf denotes where a food source is located.

Distraction implies that the target should be diverted away from a potential predator. Following is a calculation of the distraction:

$$E_i = Xe + X \tag{7}$$

where Xe denotes the enemy's location.

These five actions regulate how dragonflies migrate across DA. Each dragonfly's position is updated using the step vector generated as follows:

$$\Delta X_i(t+1) = \left(sS_i + \alpha A_i + cC_i + fF_i + eE_i\right) + \omega \Delta X_i(t+1) \tag{8}$$

where s is the weight of separation, α denotes alignment, c denotes cohesion, f denotes food weight, e denotes predator weight, t denotes the current iteration, and w denotes inertia weight.

The following equation is used to update the positions of dragonflies in the original DA:

$$X_i(t+1) = X_i(t) + \Delta X_i(t+1) \tag{9}$$

These movements and navigations enable this algorithm to address ongoing issues. In contrast to DA, BDA updates its position vectors using the following equations:

$$X_i^d(t+1) = \begin{cases} 1 - X_i^d(t) rand < TF(\Delta X_i^d(t+1)) \\ X_i^d(t) rand \geq TF(\Delta X_i^d(t+1)) \end{cases} \tag{10}$$

$$TF(\Delta X) = \left| \frac{\Delta X}{\sqrt{\Delta X^2 + 1}} \right| \tag{11}$$

where X_i^d is the ith dragonfly's dth location, *rand* denotes a number chosen at random between 0 and 1, t denotes the current iteration, ΔX denotes the step vector, and TF(.) denotes the transfer function as described in Eq. (11). BDA uses a V-shape transfer function to determine the likelihood that a dragonfly's position will change. In contrast to other binary metaheuristics, BDA does not compel the dragonflies to select values of 1 and 0. As a result, BDA has strong exploration capabilities that helped it find the relevant search space. To segment the human portion of thermal images, these extracted features are provided in the segmentation stage.

4.3 Segmentation

To segment the human in given thermal images, the extracted features from the feature extraction stage are utilized. One of the most recently developed metaheuristic algorithms is the Crow Search Algorithm (CSA), which is proposed for global optimization and used to segment the human area in input images using collected features. It is based on the crows' superior intelligence. Crows memorize where they have stashed their food so they can find it later on. To locate these hiding places and steal the food that has been stashed, they also attempt to follow one another. In these situations, the pursued crow can choose a dubious hiding place to lose its tail and fool the plunderer. To answer that question, the conventional CSA is founded on these two fundamental ideas: (a) guarding its hiding location, and (b) spotting another crow's hiding area.

4.3.1 Enhanced Crow Search Algorithm (ECSA)

Like any other metaheuristic, the Crow search algorithm might become stymied in a local optimum due to a favorable balance between exploration and exploitation. The crow and the AP parameter to follow are the major factors influencing how well it performs in terms of searchability and convergence. We explain our proposed ECSA in this section to address this weakness. We introduced a novel, improved CS

in 3 ways to improve both global and local searching: a DAP to balance intensification and diversification; an LNSS to choose which crow to follow; and a new GUPS.

4.3.1.1 Dynamic awareness probability (DAP) The primary benefit of the Crow search algorithm is that it just requires small adjustments to two parameters: awareness probability AP and flight duration FL. The value of AP regulates the ratio of exploration to exploitation and it has a significant impact on CSA's performance. The classical CSA can produce poor outcomes since the measure of AP is fixed at 0.1 for all crows since the start of the segmentation process.

In this paper, we propose utilizing a dynamic updating technique for AP to change the value of every crow's AP for every iteration depending on its rank to segment the choice of the exploitation or exploration strategy in the traditional CSA [22]. Each crow in this method has its fitness function assessed. Based on each crow's fitness value, a sorting procedure is then carried out, going from best to worst. Each crow is given a rating using Eq. (4) after this sorting.

$$rank_i = i, \quad i = 1,2, \ldots, NP \tag{12}$$

The crow with the lowest fitness, according to Eq. (12), will be awarded rank one, whereas the worst answer will be awarded the last rank.

Each member of the crow is given an AP based solely on where they fall in the individual rank, not on how fit they are. According to its rank, an AP value between $AP_{min}$and $AP_{max}$is chosen for each crow as follows:

$$DAP_i^{gen} = AP_{min} + \left(AP_{max} - AP_{min}\right)\frac{rank}{NP} \tag{13}$$

where $rank_i$ is the rank of the ith answer, and NP refers to the capacity of their population. AP_{min} and AP_{max} are the maximum and minimum values of AP. In this research, we fixed the experimental AP_{min} to 0.1 and AP_{max} to 0.8.

4.3.1.2 Improved local search Let us consider $p = \left[X_1, \ldots, X_{NP}\right]$. Every crow X is a D-dimensional parameter vector $X_i = \left[x_{1,1}, \ldots, x_{i,D}\right]$. The position updating of a crow in the original CSA is dependent on the food location of a randomly chosen crow. When the chosen crow is a poor solution, this technique may result in slow convergence.

Each crow j uses a local neighborhood selection approach to choose a crow I to follow to enhance the exploitation strategy of the traditional Crow Search Algorithm. Instead of polling the whole population as in the traditional CSA, a crow Xi employs the crow to follow from a small neighborhood to update its location. We use the assumption that the vectors are arranged concerning their indices on a ring topology, drawing inspiration from the neighborhood models presented for PSO. We establish a local neighborhood with a radius of k for each crowX_i so that its immediate neighbors are $X_{i-k}, \ldots, X_i, \ldots, X_{i+k}$.

Each crow X_i in ECSA formed a D-dimensional parameter vector $d = \left[d_1, \ldots, d_D\right]$ from its immediate surroundings, with every $d_i$representing the indices of the crow X_j to learn from its optimal location. A crow $X_j$from among its immediate neighbors is randomly chosen for each dimension once the neighborhood of the crow $X_i$has been generated. By not upgrading all of a crow X_i's dimensions from the crow X_j, we ensure better exploration and reduce the likelihood that ECSA will become stuck in a local minimum. According to Eq. (14), the crow positions have been revised.

$$x_{i,s}^{gen+1} = x_{i,s}^{gen} + FL^{gen} * \left(m_{d(j),s}^{gen} - x_{i,s}^{gen}\right); \quad s = 1 \ldots .D \tag{14}$$

Remember that the neighborhood changes as you search. After a predetermined number of generations, it is altered by randomly rearranging the crows in the population, enhancing the crows' capacity for exploration by disseminating new knowledge.

4.3.1.3 Improved global search strategy When the ith crow is aware of the presence of the jth crow in the original CSA, he purposefully follows a random track that may slow convergence. We provide a novel global method in ECSA to enhance CSA's exploration performance to prevent this problem. Following Eq. (15), we adjust the crow's position about the current best global solution so that it can investigate and make use of its surroundings.

$$if\ rand < 0.5 x_i^{gen+1} = Best + c1 * c2\ else \tag{15}$$

$$x_i^{gen+1} = Best - c1 * c2\ end$$

where Best is the overall best position and x_i is the crow's current location. According to Eq. (17), $c1$ is a component that decreases linearly across iterations, and $c2$ is a variable chosen from the range [0. 1]. A crow can therefore upgrade its location in the vicinity of the global solution using Eq. (15):

$$c_1 = 2 * exp(-4 * gen/\max\_gen)^2 \tag{16}$$

This segmented images are utilized to detect the human face in thermal images to better classification of human face emotions.

4.4 Detection

After segmentation, we utilized the Dense-UNet technique to recognize the human face in thermal face images. The four down-samplings that U-Net normally performs before

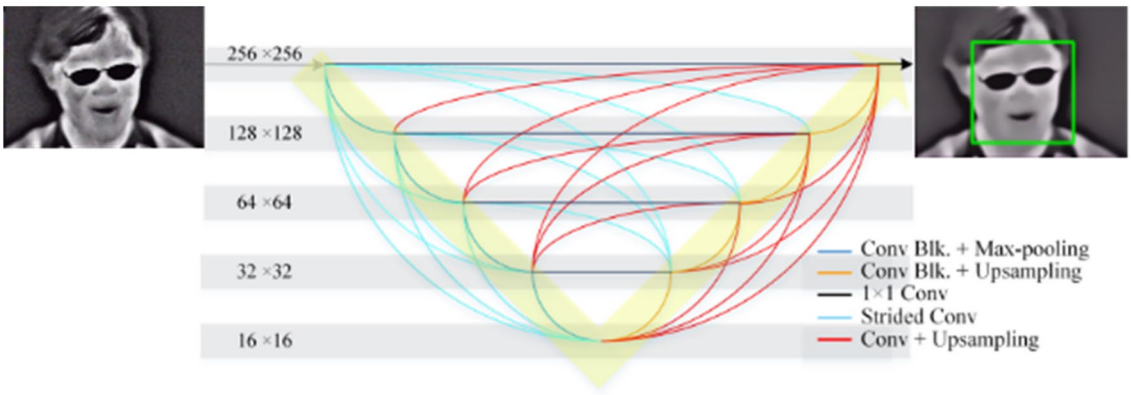

Fig. 2 The architecture of the proposed Dense-UNet model

the concatenate process result in resolution loss. Extensional procedures are required to boost accuracy because of the ensuing resolution loss. Instead of shallow network structures, these methods rely on deep ones. We chose the dense concatenated U-Net, also known as the Dense-UNet for these reasons. Our proposed Dense-main UNet argues that it can be produced by increasing data flow across the system. Convolution layers produce intermediate feature maps for CXR pictures that are remarkably similar to one another [23]. A connection design is employed to completely utilize the feature maps' capacity and prevent redundancies, which significantly reduces the cost of computing. In Dense-UNet, the outputs of numerous intermediate layers are concatenated to form the input for the following layers.

In the proposed implementation of the U-Net, which takes advantage of dense connectivity, the developed feature maps from earlier levels are utilized in all following layers (Fig. 2).

During feedforward passes, then the layers have direct access to all earlier maps. The layer has multi-level features, enabling the incorporation of various level maps. Deep supervision in the backward gradient flow makes learning easier, and gradients can expand throughout all layers, even the main ones. Convergence is facilitated by the loss function's significant influence on the model's multiple layers, and information flow enables a model with a lighter structure and significantly fewer parameters while still performing well.

The links in Fig. 2 take characteristics out of the transferred maps and balance their size. There are $\frac{9\times(9-1)}{2} = 36$ links between the 9 layers of the network. Max-pooling, Stride convolutions, and up-sampling techniques are utilized to find the issue of inconsistent sizes within various layers. The dimensions of the outcome feature maps are given in Eq. (17).

$$n_{out} = \left[\frac{n_{in} + 2p - k}{s}\right] + 1 \tag{17}$$

where n_{out} and n_{in} are the dimensions of the output and input tensors, respectively, and p indicates the padding value around the map. S is the stride step. The following details are provided for the structure shown in Fig. 2:

- The convolution block, max-pooling, and techniques used in the Dense-main UNet's body are represented by dark blue lines (down-sampling by 2). The first two components of a convolution block are two convolution layers, while the subsequent components are rectified linear activation function and batch normalization.
- Orange lines serve as a representation of the expansion path's up-sampling and convolution blocks. With these, feature maps are generated that are 2n times as big as the original map.
- The 1×1 Conv link (in black) does not change the size of the feature maps; it only alters the depth of the feature maps during feature extraction. By transferring feature maps between the output layer and the matching layers, the detection mask is finally formed.
- Smaller-scale feature maps are created by stride convolution layers, which are illustrated by the light blue lines. These connections, which can be shown in Fig. 2, transfer maps from the encoder to the decoder while tunings are used to alter the map sizes.
- The red lines, which contain an up-sampling procedure with scaling factors of 2, 4, and 8, show the final connection type.

This Dense-UNet technique detects the people's faces in thermal images to better classification of the expressions.

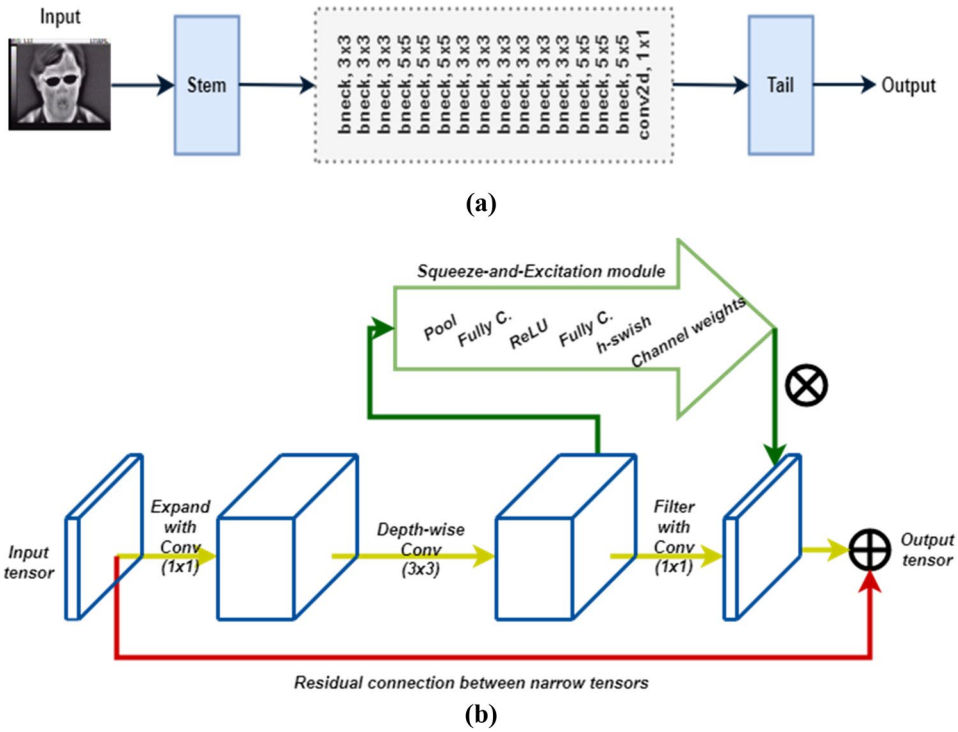

Fig. 3 A collection of bneck blocks are combined to create MobileNet functions. **a** A high-level summary. **b** Bneck block illustration

4.5 Classification

After successfully recognizing a human face in thermal images, we propose a deep learning technique of MobileNetv3 to categorize the expressions on human faces. A group of Google, Inc. researchers developed the Convolutional Neural Networks (CNN) family known as MobileNet for picture classification. Through its various iterations, MobileNet introduced a variety of cutting-edge concepts geared toward lowering the amount of parameters to make it more effective for mobile devices while retaining high classification accuracy. The inverted residual block (IRB), squeeze-excitation (SE) modules, depth-wise convolution, and a new activation function known as h-swish are some of these innovative concepts. Comparing MobileNet to many other CNN designs of comparable scale, MADDS (multiply-add operations) delivers good performance in terms of accuracy [24]. Among the many models, MobileNetV3 in particular has the highest top-1 accuracy. This is the primary driver for our decision to research the MobileNetV3 model for this classification task.

Bneck blocks are a group of building blocks that makeup MobileNet. Figure 3b depicts the specifics of a bneck block, whereas Fig. 3a depicts the broader MobileNet architecture. To minimize the number of parameters, MobileNetV1 replaced standard convolutional procedures with depth-wise convolutional operations in each block. Figure 3b shows the addition of a residual link between the output and input tensors. After that, the authors of MobileNetV2 included compression and expansion phases at the start and end of every bneck block, as depicted in Fig. 3. Due to the restricted (i.e., few channels) output and input tensor connections made by the residual connections, as opposed to the enlarged tensors found in the ResNet CNN approach, this configuration is known as an Inverted Residual Block (IRB).

The IRB concept helped to further reduce the model's computing costs. After filtering the output and input tensors, the authors utilized linear activations rather than non-linear activation functions to further minimize calculations (such as ReLU). The authors completed the MobileNetV3 model by including a SE module. In Fig. 3b, the SE module and its layers are depicted. MobileNetV3 adds the SE module concurrently with the IRB connection, in contrast to other models that add it as a distinct block of ResNet or InceptionCNN models, as illustrated in Fig. 3b. The SE module marginally grows the system's size while also enhancing its accuracy and latency.

Additionally, the authors included an h-swish activation function in the SE module. The following is a definition of the Swish activation function:

$$Swish(x) = x\sigma(\beta x) = \frac{x}{1 + e^{-\beta x}} \tag{18}$$

where β is a trainable parameter and σ(βx) is the sigmoid function. This function is referred to as the sigmoid-weighted linear unit function when β = 1. However, because it is computationally expensive to compute this function, they invented the h-swish function, which is provided by:

$$h - swish(x) = x\frac{ReLU6(x+3)}{6} \tag{19}$$

where ReLU6(x) modifies the rectified linear unit, limiting the activation to a maximum size of 6. According to the aforementioned discussion, a bneck block creates a feature map that is optimized with the help of SE modules and residual connections. This encouraged us to use it as the fundamental model of a UNet-like architecture.

5 Results and discussion

This section's first half categorizes human expressions using the dataset's evaluation and methodology for extracting human action feature attributes, comparing our method to "state-of-the-art" methods. In the next subsections, provide the assessment results based on the experimental data to evaluate our techniques.

5.1 Dataset description

5.1.1 IRIS Thermal/Visible face database

The OTCBVS benchmark data sets collection includes the IRIS thermal/visible face database. This collection includes thermal and visible face images in a range of lighting, positions, and facial expressions. It has 30 persons with a visible and thermal image size of 320 by 240 pixels in various lighting conditions. Every class contains 11 examples in the sub-database we used, which is set up with varied stances and no light.

5.1.2 IR database

The database made available by Kopaczka et al. The photographs in this database were captured with an Infratec HD820 high-resolution infrared camera with a 30-mm f/1.0 prime lens and a 1024×768 pixel-sized microbolometer sensor with a thermal resolution of 0.03 K at 30 C. The subjects were videotaped while seated at 0.9 m from the camera, outcomes in a spatial resolution of the face of roughly 0.5 millimeters per pixel. To reduce background variation, the recordings were made against a thermally neutral backdrop. The database was created by manually selecting and retrieving pictures from video footage of the participants that were taken at a frame rate of 30 frames/s. A total of 1782 sample pictures representing the eight expressions categories of disgust, anger, fear, sadness, contempt, happiness, surprise, and neutrality are present in the database.

5.2 Quantitative metrics

We proposed a method to identify facial expressions in thermal images. The noise from the provided input image is first removed using a mean filter, and the data is then normalized using the min–max normalization method. Then, the Binary Dragonfly Algorithm (BDA) approach is used to extract the features of the human face, including the Forehead, eye, chick (left and right), nose, and mouth. Then, using the Enhanced Crow Search Algorithm (ECSA), segment the human areas. Using the Dense-UNet method to identify the face in a set of thermal face photos. Finally, using the MobileNetv3 method categorizes the face into its expressions, such as neutral, sad, fearful, glad, disdain, surprise, laughing, and fury. Two separate thermal facial expression databases the IR database and the IRIS Thermal/Visible database were used in the trials to achieve good performance in terms of precision, accuracy, F1-score, and recall metrics.

The categorized expressions on faces show in Fig. 4. This is classified into eight emotions.

5.3 Evaluation metrics

The proposed approach's recall (R), precision (P), accuracy (A), and F1-score (F) were examined as performance indicators. These measurements show:

5.3.1 Accuracy

To determine whether the classification of bamboo species is accurate, the accuracy measure is calculated.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{20}$$

5.3.2 Precision

The proportion of accurately predicted positive outcomes to all predicted positive observations is known as precision. The ability to carry out the following actions is precision.

$$Precision = \frac{TP}{TP + FP} \tag{21}$$

5.3.3 Recall

The terms for the recall are sensitivity and true positive rate (TPR). The classifier's capacity to find all positive samples is shown by the recall score. It is the total divided by TP, including FN. As an example, consider the following:

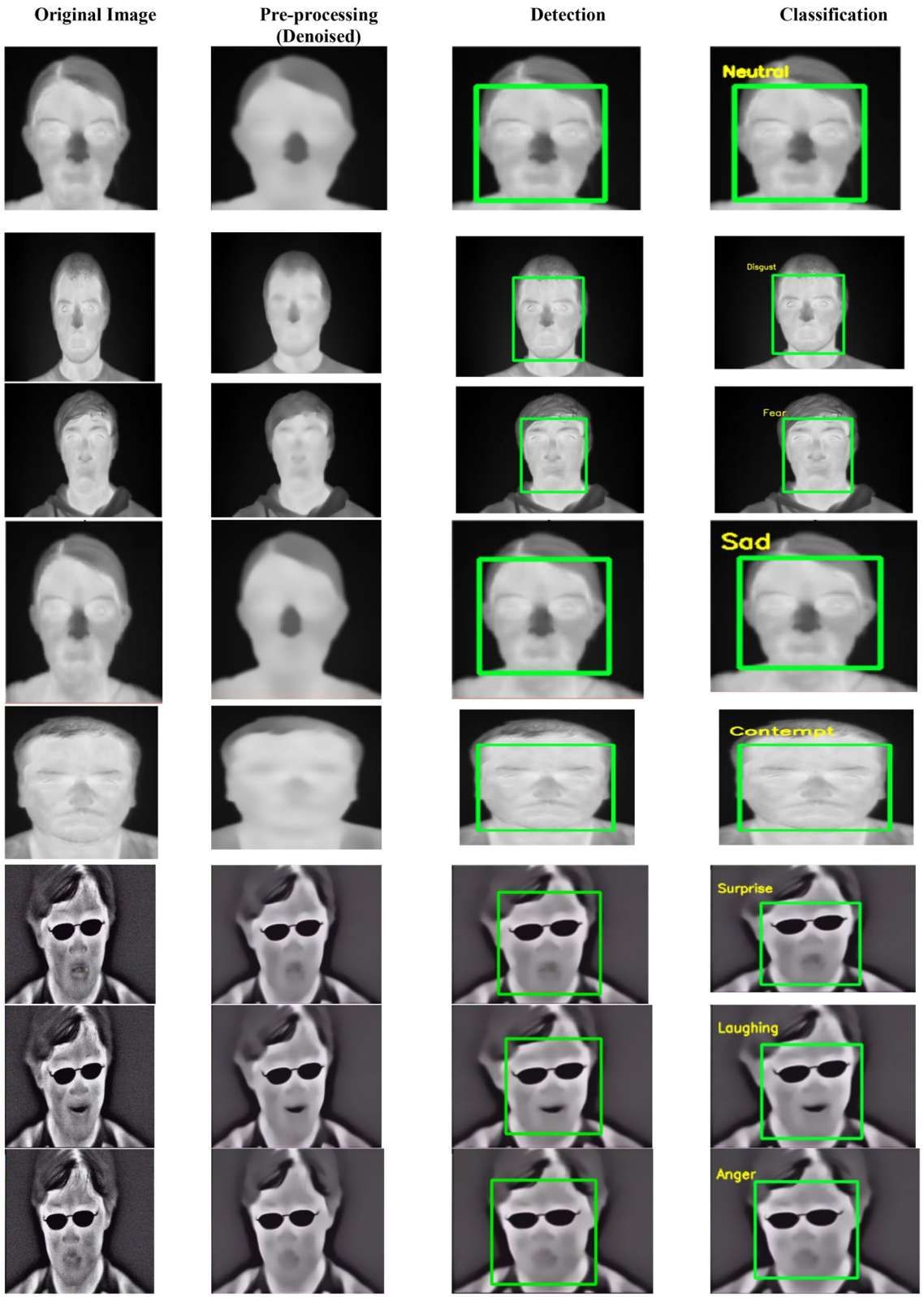

Fig. 4 The categorization results of facial emotions using the MobileNetv3 technique.

Table 1 Categorization of the proposed approach into various classes on the IR database

Emotion types	MobileNetv3 (%)			
	Accuracy	Precision	F1-score	Recall
Neutral	99.12	99.16	99.65	99.81
Disgust	99.34	99.72	99.74	99.25
Fear	99.06	99.47	99.38	99.13
Sad	99.49	99.08	99.17	99.60
Contempt	99.04	99.51	99.39	99.14

$$Recall = \frac{TP}{TP + FN} \tag{22}$$

5.3.4 F-measure

The harmonic mean of recall and precision is calculated using F-measure.

$$F1\ score = 2 \times \frac{precision \times recall}{precision + recall} \tag{23}$$

5.4 Performance evaluation

Several human expressions are done in the IR database during the execution of methods. Table 1 displays the results of the proposed method's performance measurements on the IR database. In comparison to other percentages, using the MobileNetv3 technique results in improved precision, accuracy, F1-score, and recall scores.

Figure 5 below shows how well the proposed approach performed in the IR database.

Several human expressions are done in the IR database during the execution of methods. Table 2 displays the results of the proposed method's performance measurements on the IRIS Thermal/Visible database. In comparison to other percentages, using the MobileNetv3 technique results in improved precision, accuracy, F1-score, and recall scores.

Figure 6 below shows how well the proposed approach performed in the IRIS Thermal/Visible database.

5.4.1 Comparison of IR database and IRIS Thermal/Visible database with various approaches

Performances of the IR database and IRIS Thermal/Visible database can be compared to those of the previous methods. The proposed strategy is contrasted with previous methods including DTFA, TV-CycleGAN, AGFR, and MCFG.

The proposed methodology is compared with two other methods on the IR database from Table 3. In comparison to other approaches, it achieves 98.21% accuracy and 98.69%

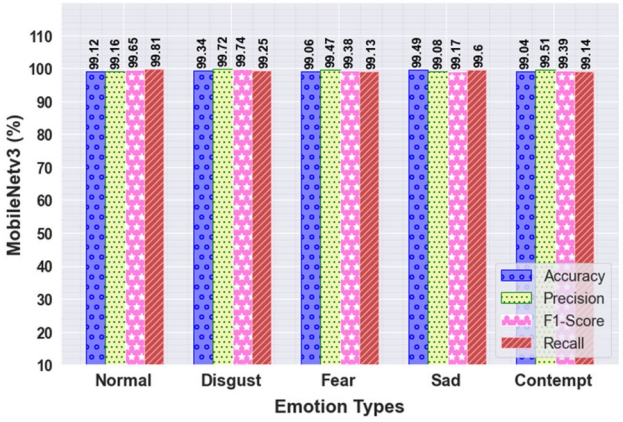

Fig. 5 The proposed approach's performance in multiple classifications in the IR database

Table 2 Categorization of the proposed approach into various classes on the IRIS Thermal/Visible database

Emotion types	MobileNetv3 (%)			
	Accuracy	Precision	F1-score	Recall
Surprise	99.40	99.25	99.74	99.05
Laughing	99.31	99.62	99.09	99.61
Anger	99.72	99.53	99.34	99.84

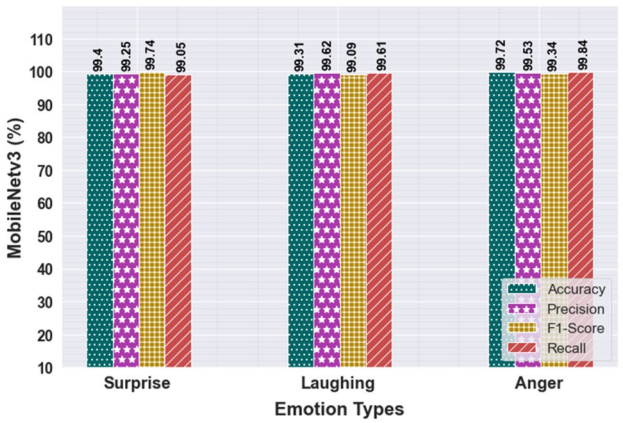

Fig. 6 The proposed approach's performance in multiple classifications in the IRIS Thermal/Visible database

TPR, and its FPR is lower at 0.76%. In comparison to the other two existing techniques, the proposed technique achieved 99.08% accuracy and 99.31% TPR in the IRIS Thermal/Visible face database. The FPR rate is lower than those of other methods now in use.

Figure 7 of the IR database presents a graphic representation of TPR and accuracy. Our proposed technique provides

Table 3 Comparison of the test result on different methods

Dataset	Approaches	Accuracy (%)	TPR (%)	FPR (%)
IR database	DTFA	89.67	89.24	0.68
	TV-CycleGAN	57.56	60.12	0.72
	Proposed	98.21	98.69	0.76
IRIS Thermal/ Visible face database	AGFR	91.03	91.55	3.96
	MCFG	94.81	94.05	5.21
	Proposed	99.08	99.31	2.51

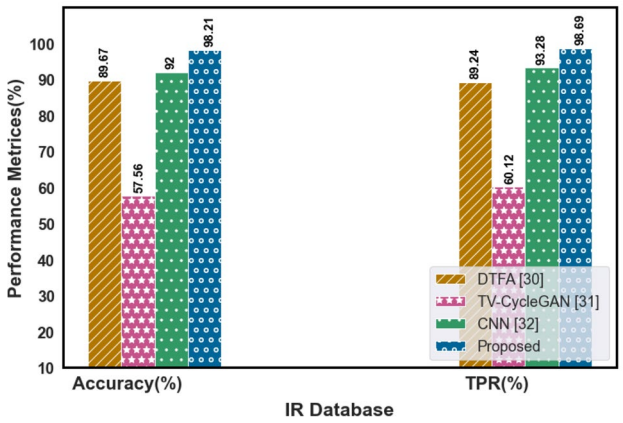

Fig. 7 Multiple performances of the IR database proposed approach with previous methods

better results when the difference can be made utilizing the previous techniques.

Figure 8 from the IRIS Thermal/Visible database illustrates TPR and accuracy. Our proposed approach gives successful outcomes when the difference can be made utilizing the previous techniques. Figure 9 compares the results of FPR measures and illustrates them.

A comparison of the proposed technique with previous techniques is shown in Table 4. It shows the performance evaluation values of accuracy, F1-score, recall, precision, FRP, and FNR.

The F1-score, recall, and precision metrics are compared to earlier methods of action detection in Table 4. A proposed approach achieved a greater accuracy of 99.78%. When compared to existing action recognition algorithms, our proposed technique performs with 99.54% precision, 99.29% recall, and 99.08% F1-score. The results show that, in comparison to existing strategies, the proposed model effectively increases the detection rate. It provide novelty to our proposed methodology and with less computational time.

The compared proposed categorization technique with previous action categorization approaches is obtain higher values, which are represented in Fig. 10 above.

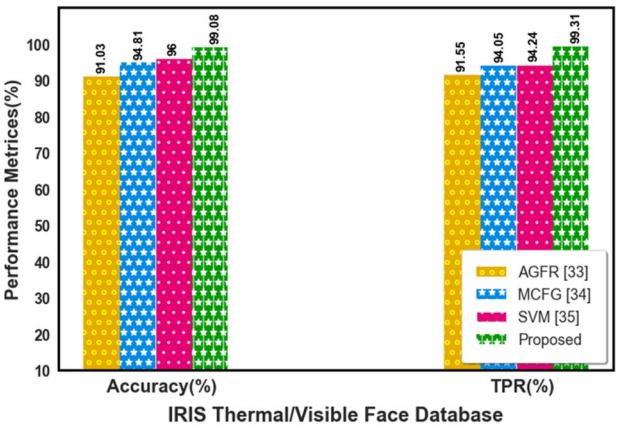

Fig. 8 Multiple performances of the IRIS Thermal/Visible database proposed approach with previous methods

Figure 11 also displays the FNR and FPR values of the proposed methods in comparison to other previous methods. Our proposed approach offers superior classification accuracy when compared to previous human action classification methods, and its testing and training values are improved.

5.4.2 Evaluation of training and testing set

Figures 12 and 13 display a graph of loss value and categorization accuracy as the number of iteration steps increased. The graph illustrates the beneficial effect of the convergence of the strategy mentioned in this study. Phases for training and testing the dataset were separated. A quarter of the testing data and seventy-five percent of the training data were created especially for this study. The proposed methods are trained for 200 iterations using the processed training set in the training phase. The learning rate is 0.1 at the moment.

The training and testing accuracy and also testing and training loss functions for two datasets are represented in Figs. 12 and 13.

6 Conclusion

Due to its consistency under different lighting conditions, thermal infrared imagery is now being used by several researchers to detect human emotion. In this research, we propose MobileNetv3, a unique deep-learning technique. To identify and categorize the emotions displayed on human faces, there are five phases. Initially, the given input image is using the mean filter to remove the noise and then normalize the data using the min–max normalization method. After that, the Binary Dragonfly Algorithm (BDA) technique is utilized to extract the features such as forehead, eye, chick

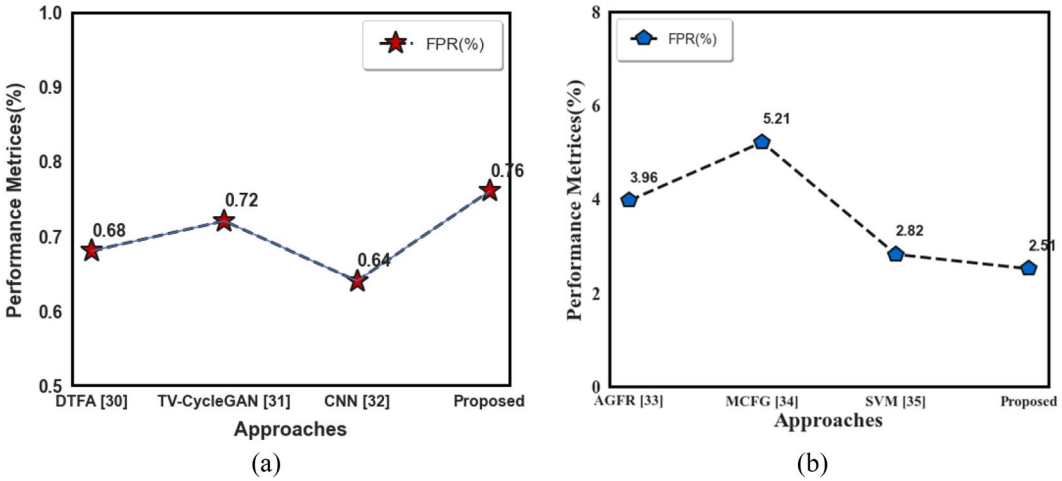

Fig. 9 The FPR performance metrics of the proposed approach with other previous techniques. **a** IR database, **b** IRIS Thermal/Visible database

Table 4 Overall performance of our proposed methodology compared with previous techniques

Approaches	Accuracy	Precision	Recall	F1-score	FRP	FNR
DTW [16]	70.59%	71.25%	70.12%	71.03%	9.74%	8.12%
IRFacExNet [17]	88.43%	88.31%	88.63%	88.07%	7.45%	9.48%
HER-BP-DWT [18]	99.55%	99.02%	99.21%	99.10%	8.32%	7.83%
FS-CNN [19]	95%	94.98%	94.39%	94.82%	9.47%	7.14%
CNN [20]	97.69%	97.24%	97.05%	97.39%	5.21%	6.30%
Proposed (MobileNetv3)	99.78%	99.54%	99.29%	99.08%	11.13%	9.23%

(left and right), nose, and mouth on a human face. Then, segment the human areas as binary segmentation using Enhanced Crow Search Algorithm (ECSA). Then, utilizing the Dense-UNet technique to recognize the face in given thermal face images. Finally, classify the face into expressions such as neutral, sad, fearful, happy, contempt, surprise, laughing, and anger utilizing the MobileNetv3 technique. The performance of experiments using two different thermal face expression databases, the IR database and IRIS

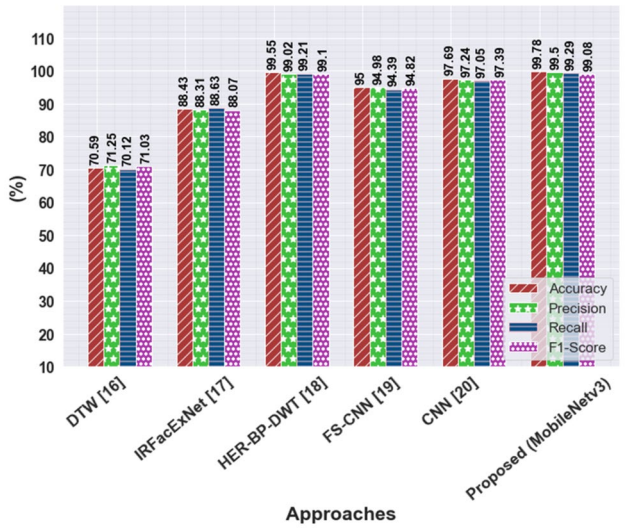

Fig. 10 Comparison of the proposed approach categorization results with previous techniques precision, accuracy, recall, and F1-score

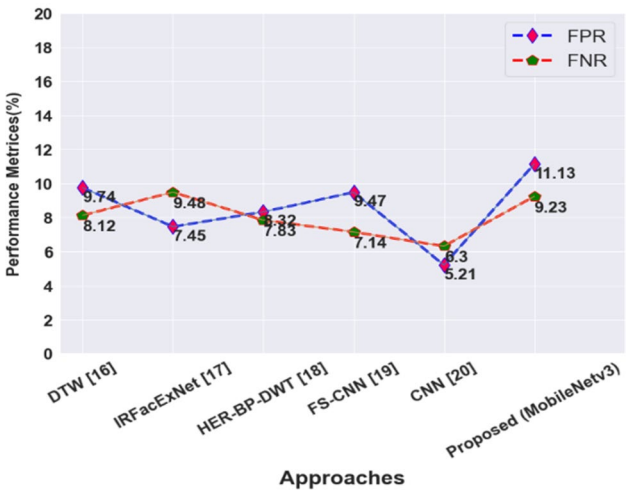

Fig. 11 Comparison of the proposed approach categorization results with existing approaches FPR and FNR

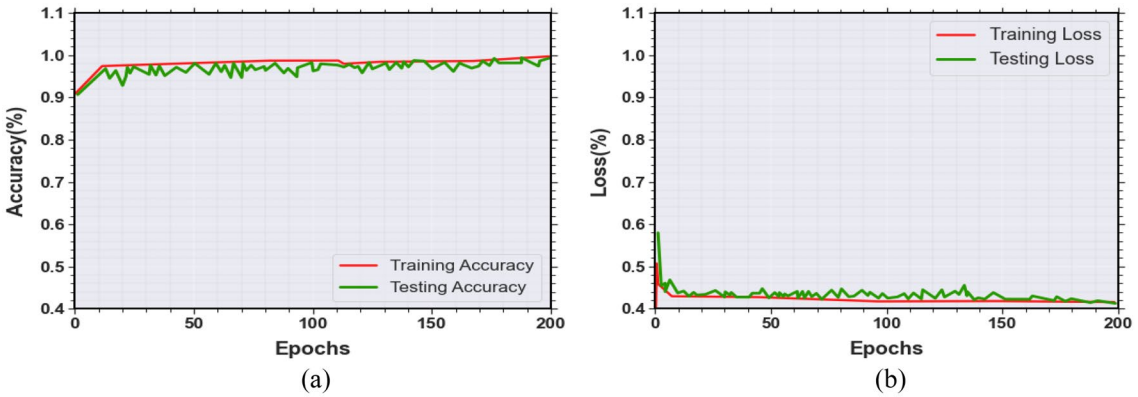

Fig. 12 **a** Training and testing accuracy, **b** training and testing loss for the IR database

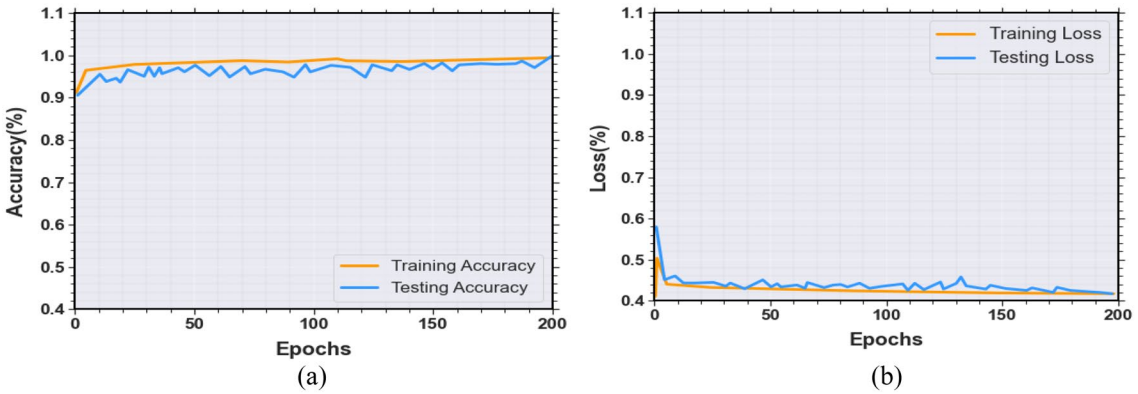

Fig. 13 **a** Training and testing accuracy, **b** training and testing loss for IRIS Thermal/Visible database

Thermal/Visible database for good performance with recall, accuracy, F1-score, and precision parameters. Our proposed method achieved the human face expression classification's accuracy is 99.78%, 99.54% for precision, 99.29% for recall, and 99.08% for F1-score. In the future, an effort might be made to create a better dataset that might potentially provide a more accurate result in day-to-day life and assist in forecasting internal security and the psychology of various people. Emotions are important in determining a person's psychology. Criminals who commit crimes in public can be detected in the case of security.

Acknowledgements We declare that this manuscript is original, has not been published before and is not currently being considered for publication elsewhere.

Author contributions The author confirms sole responsibility for the following: study conception and design, data collection, analysis and interpretation of results, and manuscript preparation.

Data availability Not applicable.

Code availability Not applicable.

Declarations

Conflict of interest 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.

Ethics approval This material is the authors' own original work, which has not been previously published elsewhere. The paper reflects the authors' own research and analysis in a truthful and complete manner.

References

1. Hossain S, Umer S, Asari V, Rout RK (2021) A unified framework of deep learning-based facial expression recognition system for diversified applications. Appl Sci 11(19):9174
2. Bashbaghi S, Granger E, Sabourin R, Parchami M (2019) Deep learning architectures for face recognition in video surveillance. In: Deep learning in object detection and recognition. Springer, Singapore, pp 133–154
3. Lee JM, An YE, Bak E, Pan S (2022) Improvement of negative emotion recognition in visible images enhanced by thermal imaging. Sustainability 14(22):15200
4. Lai YH, Chang YC, Tsai CW, Lin CH, Chen MY (2021) Data fusion analysis for attention-deficit hyperactivity disorder emotion recognition with thermal image and internet of things devices. Softw Pract Exp 51(3):595–606
5. Pons G, Ali AE, Cesar P (2020) ET-CycleGAN: generating thermal images from images in the visible spectrum for facial emotion recognition. In: Companion publication of the 2020 international conference on multimodal interaction, pp 87–91
6. Prabhakaran AK, Nair JJ, Sarath S (2021) Thermal facial expression recognition using modified resnet152. In: Advances in computing and network communications. Springer, Singapore, pp 389–396
7. Siddiqui MFH, Javaid AY (2020) A multimodal facial emotion recognition framework through the fusion of speech with visible and infrared images. Multimodal Technol Interact 4(3):46
8. Salido Ortega MG, Rodríguez LF, Gutierrez-Garcia JO (2020) Towards emotion recognition from contextual information using machine learning. J Ambient Intell Humaniz Comput 11(8):3187–3207
9. Mishra C, Bagyammal T, Parameswaran L (2021) An algorithm design for anomaly detection in thermal images. In: Innovations in electrical and electronic engineering. Springer, Singapore, pp 633–650
10. Middya AI, Nag B, Roy S (2022) Deep learning based multimodal emotion recognition using model-level fusion of audio–visual modalities. Knowl Based Syst 244:108580
11. Kezebou L, Oludare V, Panetta K, Agaian S (2020) TR-GAN: thermal to RGB face synthesis with generative adversarial network for cross-modal face recognition. In: Mobile multimedia/image processing, security, and applications 2020, vol 11399. SPIE, pp 158–168
12. Nayak S, Sharma V, Panda SK, Uttarkabat S (2021) Affective state analysis through visual and thermal image sequences. In: Emerging technologies in data mining and information security. Springer, Singapore, pp 65–73
13. Ganesh K, Umapathy S, Thanaraj Krishnan P (2021) Deep learning techniques for automated detection of autism spectrum disorder based on thermal imaging. Proc Inst Mech Eng [H] 235(10):1113–1127
14. Jiang D, Wu K, Chen D, Tu G, Zhou T, Garg A, Gao L (2020) A probability and integrated learning based classification algorithm for high-level human emotion recognition problems. Measurement 150:107049
15. Barnawi A, Chhikara P, Tekchandani R, Kumar N (2021) Artificial intelligence-enabled internet of things-based system for COVID-19 screening using aerial thermal imaging. Future Gener Comput Syst 124:119–132
16. Nayak S, Nagesh B, Routray A, Sarma M (2021) A human–computer interaction framework for emotion recognition through time-series thermal video sequences. Comput Electr Eng 93:107280
17. Bhattacharyya A, Chatterjee S, Sen S, Sinitca A, Kaplun D, Sarkar R (2021) A deep learning model for classifying human facial expressions from infrared thermal images. Sci Rep 11(1):1–17
18. Chopade PB, Prabhakar N (2021) Human emotion recognition based on block patterns of image and wavelet transform. Int J Adv Technol Eng Explor 8(83):1394
19. Said Y, Barr M (2021) Human emotion recognition based on facial expressions via deep learning on high-resolution images. Multimed Tools Appl 80(16):25241–25253
20. Umer S, Rout RK, Pero C, Nappi M (2022) Facial expression recognition with trade-offs between data augmentation and deep learning features. J Ambient Intell Humaniz Comput 13(2):721–735
21. Too J, Mirjalili S (2021) A hyper learning binary dragonfly algorithm for feature selection: a COVID-19 case study. Knowl Based Syst 212:106553
22. Ouadfel S, Abd Elaziz M (2020) Enhanced crow search algorithm for feature selection. Expert Syst Appl 159:113572
23. Cai S, Tian Y, Lui H, Zeng H, Wu Y, Chen G (2020) Dense-UNet: a novel multiphoton in vivo cellular image segmentation model based on a convolutional neural network. Quant imaging Med Surg 10(6):1275
24. Abd Elaziz M, Dahou A, Alsaleh NA, Elsheikh AH, Saba AI, Ahmadein M (2021) Boosting COVID-19 image classification using MobileNetV3 and aquila optimizer algorithm. Entropy 23(11):1383