

Original Article

YOLO-SFV2: An Effective Deep Learning Technique to Detect and Classify the Human Face Action in Thermal Images

P.R. Ajitha

Dept. of Information Technology, Dhanalakshmi Srinivasan College of Engineering and Technology, Mammallapuram, Chennai, India.

Received Date: 15 August 2023

Revised Date: 30 August 2023

Accepted Date: 15 September 2023

Abstract: Facial expression recognition (FER), a computer vision problem, tries to identify and classify the many expressions of emotion that can be detected on a person's face. One of the largest challenges to face recognizing and identification is the extraordinary variety of human faces in terms of size, shape, position, illumination, expression, and occlusion. In this research, propose a novel deep learning technique to detect and classify the human face expression in thermal images. Initially, remove noise from the input image using median filter, and then normalize the data using min-max normalization method to improve the performance. Next, use the Improved Principal Component Algorithm (IPCA) to extract the pertinent features, such as texture, shape, and location. Then, using YOLOv8 technique to detect the face expression by extracted feature. Finally, employ a new deep learning technique of ShuffleNetV2 to classify the actions into their categories. To improve the classification performance, employs the Enhanced Golden Jackal Optimization algorithm. The performance of propose methodology is evaluated on publicly available datasets are Terravic facial IR database and IRIS Thermal/Visible Face Database. State-of-the-art methodologies were used to compare the performance of the proposed approach, and the proposed approach produced higher categorization accuracy while using less processing time.

Keyword: Face Expression, Thermal Images, Shufflenetv2, Deep Learning, Classification, YOLOv8, Detection.

I. INTRODUCTION

Human action recognition is an important area of research in computer vision, with applications in gaming, robotics, autonomous observation, smart home systems, animation, and human-machine interfaces. The increase of nuclear families, the aging of the population, and nuclear families themselves have all contributed to the advancement of technology in line with supportive systems [1-3].

Several modalities of body skeletons, optical flow, and RGB images have been studied in the literature to help recognize motion in movies. Both temporal and spatial techniques are used to construct the network of two streams in human operations [4, 5]. Because human skeletons have a regular appearance and illumination, this technique has already been used to recognize human motions. Overwhelming this is the detection of action made possible by 2018Vision. However, enough lighting is necessary for cameras to function well and prevent someone from moving about unnoticed [6-8].

In many computer vision applications, predicting and recognizing human action is a key research field. It has several uses for personal protection and safety due to its help for distinguishing normal and abnormal human behavior [9, 10]. The 3D-CNN is very potent to construct both the spatial and temporal data in a short amount of time because movies of human action typically contain several frames. However, 3D Convolutional Neural Network shows incredible performance for classifying human movement since video sequences are insufficient to provide a consistent input frame [11-13].

Human action may occur in real-time and last the entire length of the film. To recognize a man's action while taking into account deep elements in our proposed framework, a unique combination of approaches is offered. minimal and high-level spatiotemporal data can be kept from the whole video sequences using this combination, which has a minimal computing cost [14, 15]. To overcome the issues in human action and expression detection and classification, propose a new deep learning technique. In this research, propose a YOLOv8 detection technique to detect the actions and expression on human face. To classify them using Shufflenetv2 approach. In pre-processing step, remove noise from the input image to better detection. It will improve the detection performance and classify it with less computation time. The key contribution of this research is,



- Initially, remove noise from the input image using median filter, and then normalize the data using min-max normalization method to improve the performance.
- Next, use the Improved Principal Component Algorithm (IPCA) to extract the pertinent features, such as texture, shape, and location. Then, using YOLOv8 technique to detect the face expression by extracted feature.
- Finally, employ a new deep learning technique of ShuffleNetV2 to classify the expressions into their categories.
- The performance of propose methodology is evaluated on publically available datasets are Terravic facial IR database and IRIS Thermal/Visible Face Database with accuracy, recall, precision, and f1-score metrics.

The rest of the research is followed in sections: Section 2 explained the previous research on human action and expression in face detection and classification. Then, the Section 3 detailed of proposed methodology. Then, the results and their discussions are mentioned in Section 4. The conclusion and future recommendation is mentioned in Section 5.

II. LITERATURE SURVEY

In this section, we mentioned and explained in detail of previous studies on human action recognition. Nan et al. [16] suggests an A-MobileNet model that is lightweight. The MobileNetV1 model first incorporates the attention module to improve the local extraction of features of facial emotions. The model variables are then optimized to decrease intra-class distance and enhance inter-class distance by combining the center loss and softmax loss. This approach considerably increases detection precision without adding more model parameters when compared to the original MobileNet series models.

A face expression detection technique based on Gabor filters and a genetic algorithm was introduced in Boughida et al. [17]. We begin by utilizing the Viola-Jones algorithm to find faces. Then, utilizing principal component analysis (PCA), reduce the Gabor features that extracted from the ROIs. The best PCA attributes are then chosen, and at the same time, utilizing GA, optimize the SVM hyperparameters to maximize the SVM model's identification rate. Finally, emotion is predicted using the SVM model.

In order to be used in face recognition applications, Litvin et al. [18] built a fully convolutional network structure for RGB picture synthesis from an input thermal face image. The suggested approach improved the FusionNet design's robustness against overfitting by utilizing orthogonal regularization, dropout after bridge connections, and randomized leaky ReLUs (RReLU). The experimental findings highlighted the advantages of using resize convolution for upscaling rather than transposed convolution and their beneficial impact when mapping thermal pictures to RGB.

To compute the multivariate time-series thermal video sequences and recognize human emotion and provide distraction ideas, Nayak et al. [19] suggested a three-stage HCI system. The first stage consists of after the face ROIs throughout the thermal video while simultaneously detecting faces, eyes, and noses utilizing a Faster R-CNN (region-based convolutional neural network) architecture. The multivariate time series (MTS) data is created by calculating the mean intensity of ROIs. The Dynamic Time Warping (DTW) technique is used to categorize the emotional states produced by video stimulus in the second stage using the smoothed MTS data. In the third stage of HCI, the suggested framework offers pertinent recommendations from a psychological and physical distraction perspective.

Said & Barr [20] proposed the use of a face-sensitive convolutional neural network (FS-CNN) to discern human emotions. Using the proposed FS-CNN, faces in large-scale pictures are identified. After that, facial landmark analysis is used to predict expression in order to identify emotions. The two stages that comprise the FS-CNN are CNN and patch cropping. Faces in high-resolution images are located and cropped in the first stage so they can be processed further. In the second phase, a CNN was used to forecast facial expressions based on landmark analytics and process scale invariance on pyramid pictures.

From the above-mentioned literatures, they have some problems such as large-sized databases which would help us to ensure the robustness technologies, the cost of time and the complexity of space are high. To overcome these issues, propose a novel deep learning techniques in this research.

III. PROPOSED METHODOLOGY

Recognizing and categorizing human behaviors and emotions from video or picture sequences is a task for computer vision and DL known as "human action identification." This study presents a fresh deep learning method to address the shortcomings of earlier research. Initially, remove noise from the input image using median filter, and then normalize the image

using min-max normalization method to improve the performance. Next, use the Improved Principal Component Algorithm (IPCA) to extract the pertinent features, such as texture, shape, and location. Then, using YOLOv8 technique to detect the face expression by extracted feature. Finally, employ a new deep learning technique of ShuffleNetV2 to classify the expressions into their categories. To improve the classification accuracy using Enhanced Golden Jackal Optimization (EGJO) algorithm. Figure 1 shows the architecture of the suggested methodology.

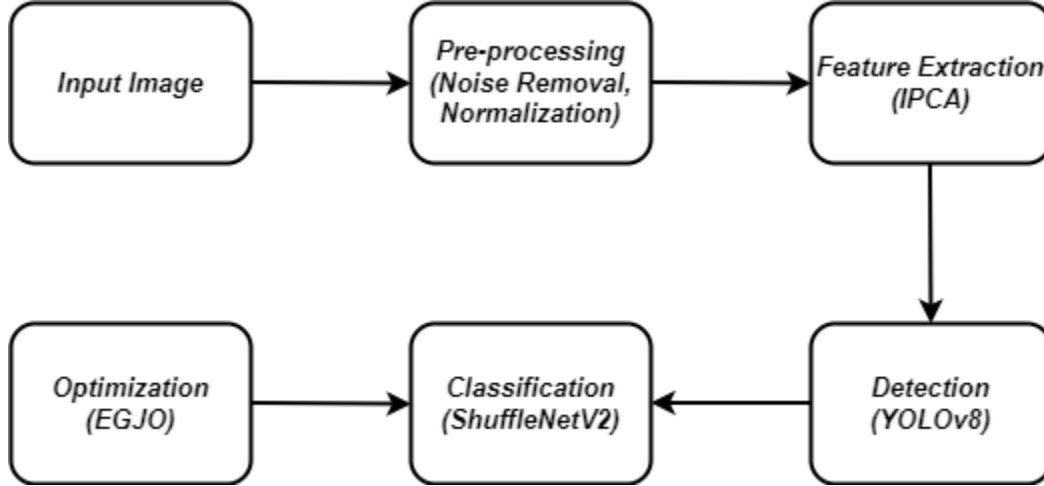


Figure 1: Architecture of Proposed Methodology

A. Earlier Handling:

Use the appropriate techniques to remove noise during the pre-processing step and to do performance normalization. The median filter, which employs a weighted average sum of the adjacent pixels, removes such noise. When it comes to keeping an image's edges, the median filter performs admirably. Images are run via a median filter after noise removal. When using a large amount of data, normalization is a handy method for scaling the image so that they fit within a specific range. Gradient descent is accelerated and becomes more precise after normalization. By applying a linear modification to the initial data, min-max normalization is frequently used to scale the data between particular ranges. The notations x_{min} and x_{max} , respectively, signify an attribute's lowest and greatest values. Calculation is used to ascertain the distinction between the two values after mapping the value x to a value within the range[x_{min} and max_x].

$$z_n = \frac{x - x_{min}}{x_{max} - x_{min}} (New_{max_x} - New_{min_x}) + New_{min_x} \tag{1}$$

Where the x_{max} and x_{min} variables, respectively, reflect the maximum and lowest possible values. The notation New_{min_x} represents the lowest amount, whereas the highest amount is represented by the notation New_{min_x} .

B. Feature Extraction:

Use the Improved Principal Component Analysis (IPCA) technique to extract features such as texture, shape, and location once pre-processing is complete. The approach of data organization and extraction of features known as PCA is efficient. In recognition of patterns, image processing, and computer vision, it attracted a ton of interest. The goal of PCA is to determine the optimum vector space that best depicts the distribution of face pictures and to reduce a huge data set to a smaller dimension in order to achieve effective results. PCA creates linear combinations of the original data. The dimension of the original space is significantly reduced by the feature space created by the eigenvector, which decreases the computation time for detecting faces and identification. The primary objective of the PCA method is to shrink the enormous face data dimensions to those of the smallest spaces. A multivariate analysis technique based on eigenvectors is known as PCA. There are two primary ways to implement the PCA algorithm. The first one is accomplished through the division of the eigenvalue of the data covariance matrix, whilst the second one is accomplished through the decomposition of a single data matrix value. The component or factor scores

and normalized component score weight is used to express PCA results. The value of every eigenface to which the generated image was associated can therefore be stated as the outcome. The visuals produced when the data set's dimensions are reduced are referred to as eigenfaces (or eigenvectors). Each pixel is taken into account as an individual dimension in a picture using the PCA eigenfaces approach. The enhanced PCA algorithm substitutes the mean of each class for the individual picture within the class as opposed to the conventional PCA. The average of each class preserves a significant number of variants of the particular image because it is a linear mixture of within-class pictures. In other words, recognizing images is made easier by the compression procedure used for every image. Additionally, the training time is much decreased, which is another clear benefit of the improved PCA.

Where m stands for picture pixels, the gray picture of the k -th input face is displayed by. First, the average values of N pictures are calculated using the formula $\bar{x} = \frac{1}{N} \sum_{i=1}^N x_i$; Second, the vector covariance matrix is calculated:

$$W = \frac{1}{N} \sum_{x=1}^N (x_i - \bar{x})(x_i - \bar{x})^T \quad (2)$$

The entire dispersion matrix is another name for this covariance matrix. It is divided into two sections: scatter matrices between classes and discrete matrices within classes. Only the between-class scatter matrix is calculated by IPCA [21]; the intra-class scatter matrix is not taken into account. In picture collections, there are N pieces of pictures that are separated into c classes. N_i is the quantity of the i -th subject, and x_i stands for the row vector of the i -th face image. Each class's average value is: $\bar{x} = \frac{1}{N_i} \sum_{i=1}^{N_i} x_i$. The standard average values are $\bar{x} = \frac{1}{N} \sum_{i=1}^N x_i$, is:

$$S = \frac{1}{N_i} \sum_{i=1}^c (x_i - \bar{x})(x_i - \bar{x})^T x_k = [x_{k1}, x_{k2}, \dots, x_{kn}]^T \quad (3)$$

S is a symmetric matrix, and its diagonalization is as follows:

$$S = \sum_{r=1}^N \lambda_r V_r V_r^T = V \Lambda V^T \quad (4)$$

The projection of x_k on V_r is denoted as: where λ_r is the eigenvalue of W , V_r is the appropriate eigenvector, $\{V_1, V_2, \dots, V_n\}$ is the standard orthogonal basis, R is the rank of W , Λ is diagonal matrix, and eigenvalues w are ordered at the diagonal.

$$P_k^T = V_r^T x \quad (5)$$

$$Var(P_k) = E[(V_r^T x_k - V_r^T \bar{x}) \cdot (V_r^T x_k - V_r^T \bar{x})^T] \quad (6)$$

$$= V_r^T E[(x_k - \bar{x}) \cdot (x_k - \bar{x})^T] V_r \quad (7)$$

$$= V_r^T W V_r = \lambda_r \quad (8)$$

C. Detection:

Finding and recognizing every visible face in a single image or video—regardless of its size, age, orientation, or expression—is the task of face recognition. Additionally, the location needs to be suitable for incidental lighting as well as the image and video material. YOLO is now the most widely used real-time object detector due to its effective feature fusion algorithms, lightweight network design, and better detection results.

YOLOv8 [22] was launched to merge the best characteristics of various real-time object detectors. The idea of CSP was still present in YOLOv5, the approach to feature fusion (PANFPN), and the SPPF module. The important developments were as follows: It provided an entirely new SOTA paradigm with the P5 640 and P6 1280 resolution object recognition networks and the YOLACT instance segmentation methodology. In order to meet the needs of diverse applications, it also developed algorithms at different scales based on a scaling coefficient similar to YOLOv5. (b) Using the YOLOv7 ELAN structure, the C2f component was created with the premise that the YOLOv5 concept would remain unchanged. (c) The detection head section also used the separation of detecting and categorizing heads. The majority of the added elements remained based on the basic idea of YOLOv5. (d) The YOLOv8 classification loss was treated as a BCE loss. VFL recommended a reduction Asymmetric weighting for a loss of

the kind CIOU loss + DFL. DFL: The location of the box was modelled using a generic distribution. Equation (1) demonstrates that the probability density was as close to the site as was practical, and the network focused right away on its distribution of where it was nearest to the object location. S_i is the network's sigmoid results, y is a label, and y_i and y_{i+1} are interval orders. When compared with the original YOLO algorithm, YOLOv8 is a more extensible approach. It's a framework that can handle and transition between previous YOLO versions, which makes assessing each one's effectiveness easier. Equation (1) indicates that the probability density was as close to the site as was practical, and the network concentrated on the distribution of this density's proximity to the object position right away. S_i is the sigmoid output of the network, y is a label, and y_i and y_{i+1} are interval orders. By comparison, YOLOv8 is a more extensible approach than the original YOLO algorithm. It is a system that can manage and switch between previous iterations of YOLO, making it easy to evaluate each version's performance.

$$DFL_{(S_i, S_{i+1})} = -(y_{i+1} - y) \log(S_i) + (y - y_i) \log(S_{i+1}) \quad (9)$$

Anchor-Free is used by YOLOv8 in place of Anchor-Base. V8's matching algorithm was Dynamic TaskAlignedAssigner. Equation (2), where s is the classification score, u is the IOU value, and a and b are the weight variables, is used to determine the alignment degree of the Anchor-level for each instance. Next, it trains the loss function using the m anchors that have the highest value (t) in each instance as positive samples and the remaining anchors as negative samples. After the aforementioned improvements, YOLOv8, the most accurate detector to date, is 1% more accurate than YOLOv5.

$$t = s^a \times u^b \quad (10)$$

The main strength of YOLOv8 is its extensibility. Academics working on YOLO projects may easily compare the outcomes because YOLOv8 is made to work with all versions of YOLO and transition between them. The YOLOv8 baseline version was chosen as a result. The C2f module, which is based on the CSP principle, replaces the C3 module, while the core of YOLOv8 is substantially identical to that of YOLOv5. By combining C3 with ELAN and expanding on the ELAN idea from YOLOv7, the C2f module allowed YOLOv8 to accept more gradient flow data while still remaining portable. The most widely used SPPF module was still in use at the backbone's end, and three 5x5 Max-pools were transmitted serially before each layer was concatenated to ensure light weight and accuracy of objects at various scales. YOLOv8 is still using the PAN-FPN feature fusion technique in the neck region, which strengthens the fusion and exploitation of features layer data at multiple scales. The neck module was created by the YOLOv8 authors using the final decoupled head framework, several C2f components, and two up-sampling. YOLOv8's final neck component was constructed using the same process as YOLOx's head. Accuracy was increased by combining regression boxes and confidence.

All YOLO versions can be supported by YOLOv8 and switched between at any time. Its capability to function on a number of hardware platforms (CPU-GPU) accounts for its considerable adaptability.

D. Classification:

The categorization procedure receives the human face that was detected as input. Utilize the ShuffleNetV2 deep learning approach to categorize person behavior and expression. ShuffleNetV2 [23], a lightweight CNN framework that's used as the model training architecture. This model was chosen because, according to impression, it integrates feature reusability with little training data. With cutting-edge performance metrics and evaluation assessment parameters, this decreases model complexity and training time.

There were 41 classes, and parameters like pooling, scale factor, the amount of shuffle units, and bottleneck ratio were set. To acquire the best results from the model, the amount of hidden layers was reduced utilizing custom inputs and hyperparameter optimization. Six building blocks, each comprising two convolutional layers, were stacked to produce the network. During the execution of the model, no layers were frozen. The 'ReLU' function was applied to each activation layer, and then a final 'softmax' layer with 41 neurons—corresponding to the number of output classes—was added. The error rate was decreased using the optimizer, Adam. The following is a list of the update weights' formulae.

The beginning of the weights

$$\rho_m \leftarrow 1, \rho_v \leftarrow 1, m \leftarrow 0, v \leftarrow 0 \quad (11)$$

Update rules for Adam Optimizer:

$$\rho_m \leftarrow \beta_m \rho_m \quad (12)$$

$$\rho_v \leftarrow \beta_v \rho_v \quad (13)$$

$$m \leftarrow \beta_m m + (1 - \beta_m) \nabla_w J \quad (14)$$

$$v \leftarrow \beta_v v + (1 - \beta_v) (\nabla_w J \odot \nabla_w J) \quad (15)$$

$$m \leftarrow w - \alpha \left(\frac{m}{\sqrt{v} + \varepsilon} \frac{\sqrt{1 - \rho_v}}{1 - \rho_m} \right) \quad (16)$$

where m, v are, respectively, the first and second moment vectors. Similar to this, β_m , and β_v reflect the first and second moment vectors' respective exponential decay rates. ρ_m and ρ_v details the temporal decay factor for the adaptive learning rate. This variable, which is related to the memory for previous weight adjustments, is comparable to momentum. α Reflects learning rate or step size in Eq. 16. $\nabla_w J$ represents the gradient of the cost function and $J \cdot \varepsilon$ is modest in Eq. 16 to avoid the division by zero constraint. Initial weights prior to update are shown by Eq. 11, and conditions that obey value and biased change are indicated by Eqs. 12 to 15. Eq. 16 displays the final modification to the weight variable. Eq. 15 uses element-wise multiplication, as well while Eq. 16 uses element-wise operation handling for operations under the root.

AdaGrad with circumstances and RMSProp are combined to create Adam optimizer. As a result, the primary moving average is transformed into Nesterov accelerated momentum, radically merging to global minima in any case train time. With its updated weights, Adam outperformed all other optimizers in the given dataset. The Enhanced Golden Jackal Optimization (EGJO) algorithm was used in this study to enhance classification performance.

E. Optimization:

A unique optimization method, the fundamental GJO was inspired by the cooperative attacking strategy of golden jackals. Every golden jackal in GJO stands for a potential solution or search engine.

a) Search Space Formulation:

To get an evenly distributed candidate answer in the search region, start the random prey population in GJO. The initial answer is calculated as:

$$Y_o = Y_{min} + rand(Y_{max} \cdot Y_{min}) \quad (17)$$

Whereas rand indicates a uniform arbitrary vector in the range [0, 1], Y_o expresses the location of the initial population of golden jackals, and Y_{min} and Y_{max} express the solution's lower and upper bounds.

b) Exploration Phase or Searching the Prey:

Although the prey occasionally swiftly evades and escapes the jackals' foraging, the golden jackals are able for predicting and capture the prey in accordance with their own attacking features. In order to wait and look for alternative prey in the search area, the female jackals follow the male jackals. The positions are calculated as follows:

$$Y_1(t) = Y_M(t) - E \cdot |Y_M(t) - rl \cdot Prey(t)| \quad (18)$$

$$Y_2(t) = Y_{FM}(t) - E \cdot |Y_{FM}(t) - rl \cdot Prey(t)| \quad (19)$$

$Y_M(t)$ and $Y_{FM}(t)$ reflect the present locations of the female and male jackals, respectively, while $Prey(t)$ expresses the location vector. The updated locations of the female and male jackals are expressed by $Y_1(t)$ and $Y_2(t)$.

The evading energy of prey E is computed as:

$$E = E_1 * E_0 \tag{20}$$

E_0 shows the initial condition of the energy, while E_1 expresses the prey's diminishing energy.

$$E_0 = 2 * r - 1 \tag{21}$$

Where r expresses an arbitrary value in [0,1].

$$E_1 = c_1 * (1 - (t/T)) \tag{22}$$

Where T denotes the highest iteration, c_1 denotes a constant with a value of 1.5, and E_1 deviates linearly from 1.5 to 0 depending on the number of iterations.

Based on the Levy distribution, the rl expresses any arbitrary vector and is calculated as follows:

$$rl = 0.05 * LF(y) \tag{23}$$

c) *Exploitation Phase or Enclosing and Pouncing the Prey:*

When the jackal pair attacks, the animal's ability to evade will quickly diminish. The golden jackals then quickly enclose and seize the prey. The positions are calculated as follows:

$$Y_1(t) = Y_M(t) - E * |rl * Y_M(t) - rl * Prey(t)| \tag{24}$$

$$Y_2(t) = Y_{FM}(t) - E * |rl * Y_{FM}(t) - rl * Prey(t)| \tag{25}$$

$Y_M(t)$ and $Y_{FM}(t)$ reflect the current locations of the female and male jackals, respectively, while $Prey(t)$ provides the location vector. The updated locations of the female and male jackals are expressed by $Y_1(t)$ and $Y_2(t)$. The previous section gave some parameters, including E and rl. The final step is to compute formula (25), which updates the golden jackal's position.

d) *Enhanced Golden Jackal Optimization (EGJO)*

The basic GJO is enhanced with the elite opposition-based technique for learning and the simplex method to overcome the drawbacks of poor optimization effectiveness and slow search speed [24]. This increases the convergence rate and improves computation precision.

i) *Elite Opposition-Based Learning Technique:*

The effective and reliable elite opposition-based method of learning broadens the search area, promotes population diversity, prevents premature convergence, and intensifies the global search. The search method makes use of the feasible or reverse solution to evaluate the fitness value of prey before selecting the most suitable candidate to carry out the iteration. Assuming that the search agent with the highest fitness value is regarded as an expert individual, the viable solution is computed as $x_i = (x_{i,1}, x_{i,2}, \dots, x_{i,D})$ and the elite individual is computed as $x_j = (x_{j,1}, x_{j,2}, \dots, x_{j,D})$. The calculation is as follows:

$$x_{i,j} = k * (da_j + db_j) - x_{c,j}, \quad i = 1, 2, \dots, n; j = 1, 2, \dots, D \tag{26}$$

Where n is the population size, D is the issue's dimension, and k is an arbitrary number, then $k \in (0,1)$, da_j and db_j correspondingly express the dynamic limits of the jth decision variable. These values are computed as follows:

$$da_j = \min(x_{i,j}), db_j = \max(x_{i,j}) \tag{27}$$

The dynamic limit can modify the search region for the inverse answer and save the best solution. Computing the search agent $x_{i,j}$ results in:

$$x_{i,j} = rand(da_j, db_j), \text{ if } x_{i,j} < da_j \text{ or } x_{i,j} > da_j \tag{28}$$

ii) *Simplex Technique:*

The simplex method is an important and useful method that speeds up the search process, enhances computational accuracy, deepens optimization, and strengthens local search. By comparing the workable solution to the starting solution, the simplex method keeps the best answer.

IV. RESULTS AND DISCUSSION

This section's first section compares our approach to cutting-edge techniques by categorizing human facial activity and emotion using an examination of the dataset and our technique for retrieving face features. Give the assessment findings based on the experimental information in the following subsections to assess our approach.

A. Dataset Description:

For the evaluation, chose two human face action and expressions related dataset. They are,

a) *Terravic Facial IR Database:*

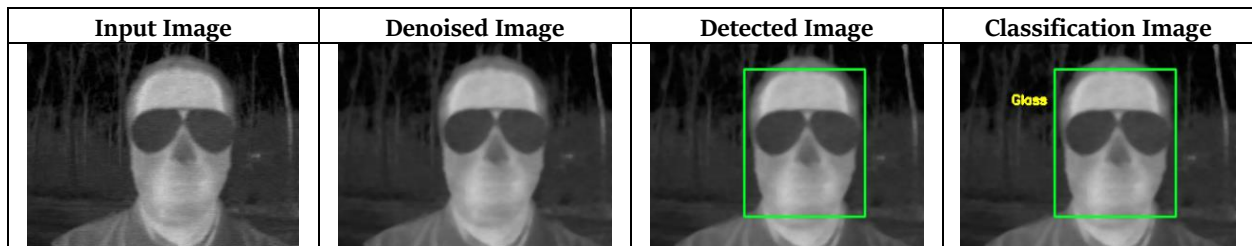
The effectiveness of this tactic was evaluated using the Terravic Facial IR Database. Every of the 20 classes in the dataset collection is represented by a set of greyscale pictures (360 240), one for each class. We utilized 17 classes in this paper because three of them were corrupted. For each class, 200 greyscale photos in total were used. The pictures are 320 x 240 pixel 8-bit grayscale JPEG files.

b) *IRIS Thermal/Visible Face Database*

The IRIS thermal/visible face database is part of the collection of benchmark data sets for OTCBVS. Both visible and thermal facial images in various lighting conditions, poses, and expressions are included in this collection. It features 30 people in 320 by 240 pixel visible and infrared images under different lighting situations. The sub-database we used has 11 samples for each class, each of which is configured with different stances and no light.

B. Implementation Results

The performance of proposed method results is shown in Figure 2. First, input thermal image is given. Then, remove noise from the input images and normalize the images. After that, extract the features using IPCA technique. Then, detect the face part based on extracted features. Finally, classify the face expressions and actions into their category.



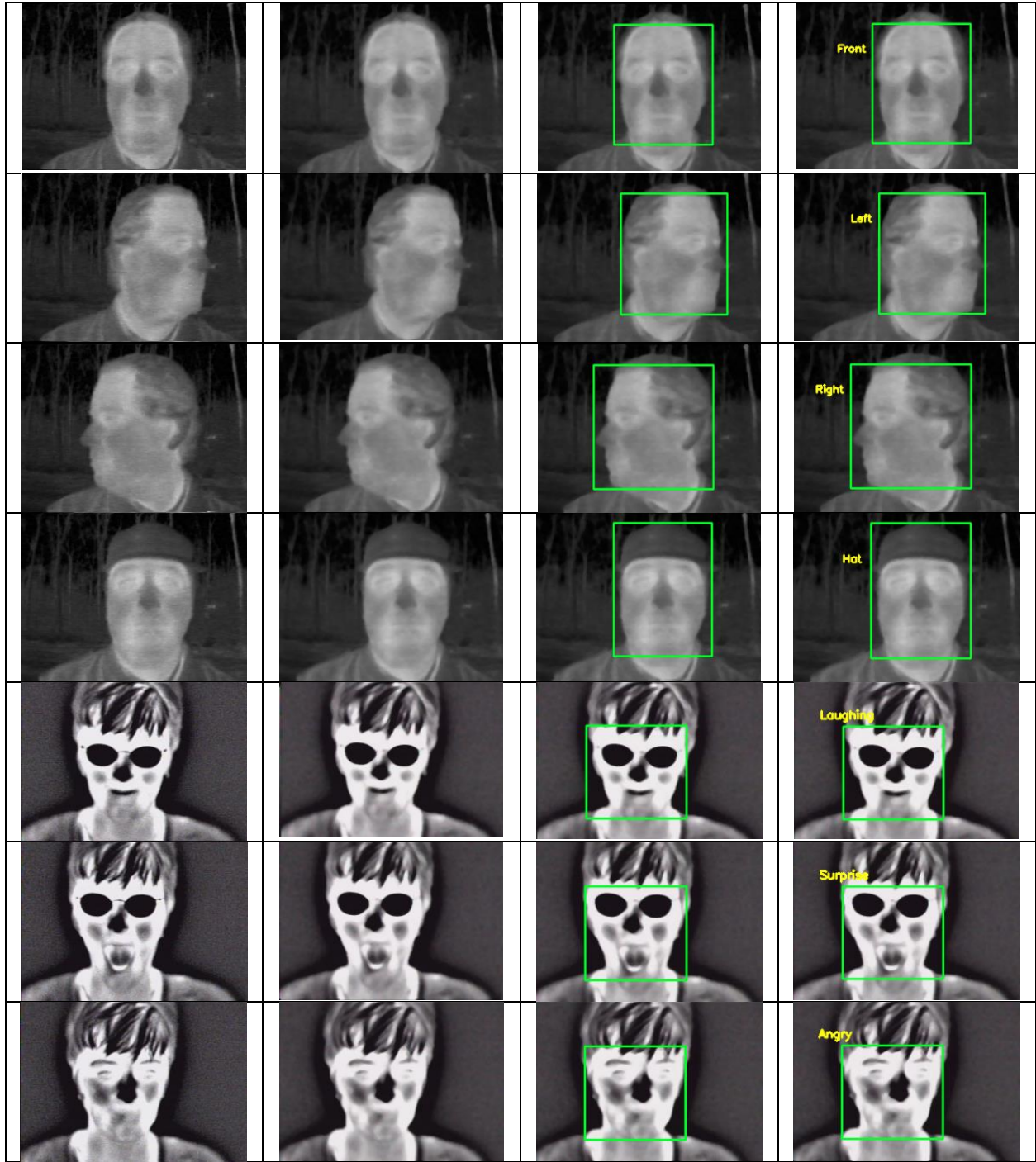


Figure 2: A result of Proposed Methodology

C. Evaluation Metrics:

The proposed technique's Precision (P), Accuracy (A), Recall (R), and F1-score (F) were examined as performance indicators. These measurements show:

a) Accuracy:

Accuracy is defined as the proportion of accurately identified samples to all samples. Generally speaking, the higher the accuracy, the better the classifier. The definition of accuracy is given in Equation (29).

$$Accuracy = \frac{TP+TN}{TP+FN+FP+TN} \quad (29)$$

b) *Sensitivity:*

Sensitivity, also known as recall, measures how well a classifier can identify positive samples by representing the percentage of all positive samples that are correctly identified. In Eq. (30), the sensitivity is described.

$$Sensitivity = \frac{TP}{TP+FN} \quad (30)$$

c) *Specificity:*

Specificity measures the classifier's capacity to identify negative samples by representing the percentage of all negative samples that are successfully classified. In Eq. (31) is a definition of specificity.

$$Specificity = \frac{TN}{TN+FP} \quad (31)$$

d) *Precision:*

Precision is defined as the ratio of correctly predicted positive outcomes to all correctly predicted positive observations. Precision is the capacity to perform the following tasks.

$$Precision = \frac{TP}{TP+FP} \quad (32)$$

D. Performance of Proposed Methodology:

The evaluation performance of the proposed methodology is shown as graphs in these sub-sections. The proposed method with two thermal face and action datasets compared with existing recognition techniques and calculated the computation time for proposed method and also compared with existing methods. While compare to other approaches, the proposed method achieved higher classification accuracy with less running time.

a) *Evaluation Performance on Dataset 1:*

Here, the proposed method obtained the values with four metrics for each class in dataset 1 are shown in Table 1. Each class achieved different level of accuracy, precision, recall, and f1-score values using proposed method.

Table 1: Multiple Class Classification using Proposed Methodology in Dataset 1

Classes	Precision	Recall	Accuracy	F1-score
Glass	99.23	99.41	99.54	99.46
Front	99.72	99.47	99.27	99.68
Left	99.31	99.69	99.48	99.52
Right	99.63	99.38	99.29	99.42
Hat	99.64	99.25	99.19	99.53

A Figure 2 shows the multi-class classification accuracy, recall, precision, and f1-score values obtained by proposed methodology in dataset 1.

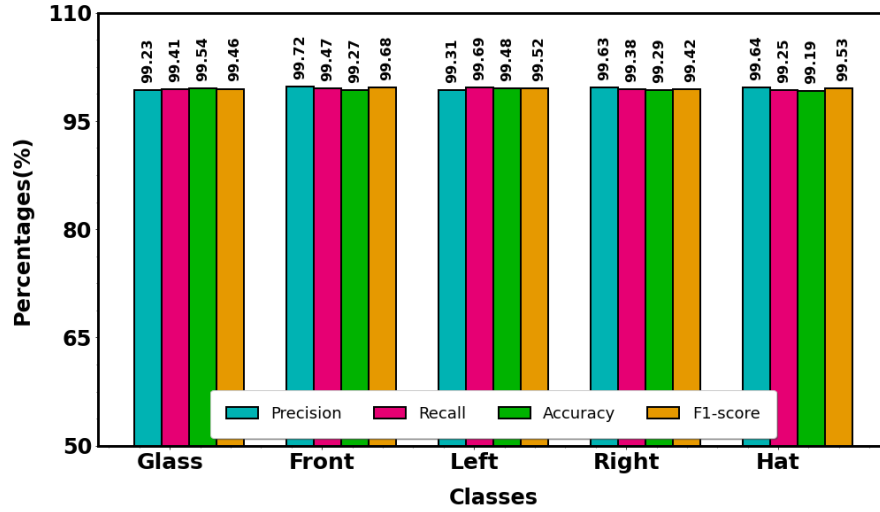


Figure 2: The Proposed Method’s Multi-Class Classification Results on Dataset 1

And Table 2 demonstrates the comparison of the proposed method with existing methods performs on dataset 1.

Table 2: Comparison of Proposed Method with other Techniques used in Dataset 1

Techniques	Accuracy	Recall	Precision	F1-score
RLO	94.12	-	-	-
MLPNN	89.5	-	-	-
HOG-SVM	98.43	-	-	-
Proposed Method	99.45	99.37	99.48	99.62

A Figure 3 displays the comparison of proposed method and existing methods perform on dataset 1.

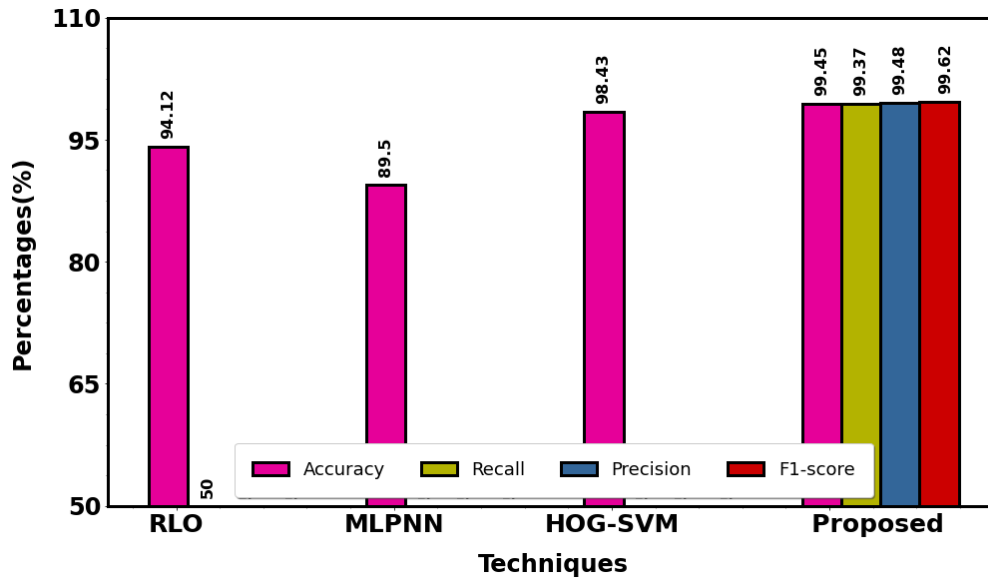


Figure 3: The Comparison of Proposed and Existing Methods on Dataset 1

b) Evaluation Performance on Dataset 2:

Here, the proposed method obtained the values with four metrics for each class in dataset 1 are shown in Table 3. Each class achieved different level of accuracy, precision, recall, and f1-score values using proposed method.

Table 3: Multiple Class Classification Using Proposed Methodology in Dataset 2

Classes	Precision	Recall	Accuracy	F1-score
Laughing	99.57	99.28	99.47	99.51
Surprise	99.64	99.49	99.53	99.45
Angry	99.24	99.52	99.46	99.59

A Figure 4 shows the multi-class classification accuracy, recall, precision, and f1-score values obtained by proposed methodology in dataset 2.

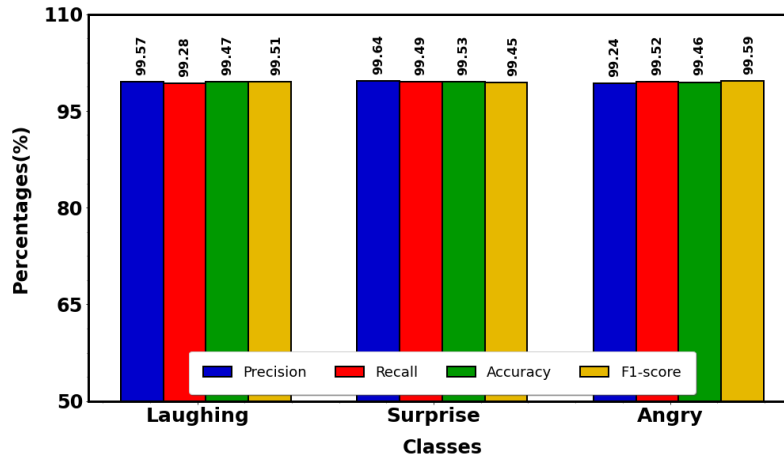


Figure 4: The Proposed Method’s Multi-Class Classification Results on Dataset 2

And Table 4 demonstrates the comparison of the proposed method with existing methods performs on dataset 2.

Table 4: Comparison of Proposed Method with Other Techniques Used In Dataset 2

Techniques	Accuracy	Recall	Precision	F1-score
SVM	97.2	-	-	-
ANN	85.3	-	-	-
Proposed Method	99.53	99.37	99.62	99.58

A Figure 5 displays the comparison of proposed method and existing methods perform on dataset 2.

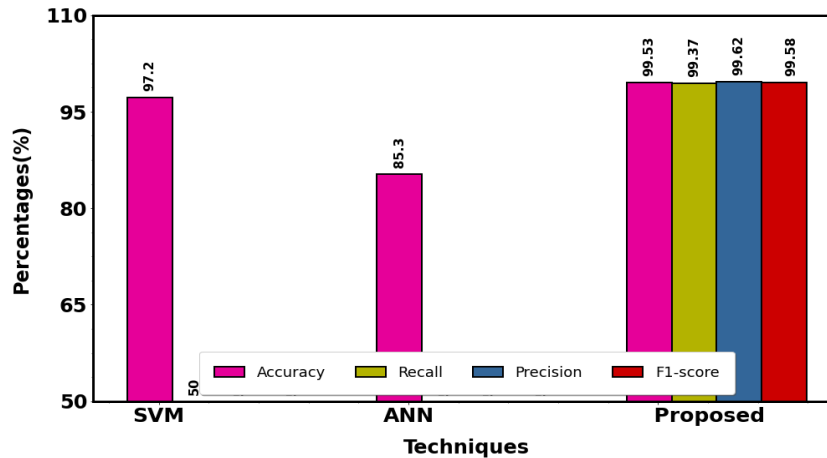


Figure 5: The Comparison of Proposed and Existing Methods on Dataset 2

c) Overall Comparison of Proposed Methodology:

The overall comparison of proposed method and other existing methods are illustrated in Table 5. This table shows the proposed method of Optimized ShuffleNetV2 outcomes with previous techniques.

Table 5: Comparison Outcome of Proposed and Existing Methods

References	Techniques	Accuracy
Nan et al. [16]	A-MobileNet	88.11
Boughida et al. [17]	SVM	98.15
Litvin et al. [18]	FusionNet	86.94
Nayak et al. [19]	DTW	70.59
Said & Barr [20]	FS-CNN	94.9
Proposed Method	Optimized ShuffleNetV2	99.56

When compared to other techniques, the proposed method achieved higher classification accuracy is 99.56% than 88.11% for A-MobileNet, 98.15% for SVM, 86.94% for FusionNet, 70.59% for DTW, 94.9% for FS-CNN. Compared to the other techniques, the proposed method achieved higher classification accuracy with less computation time.

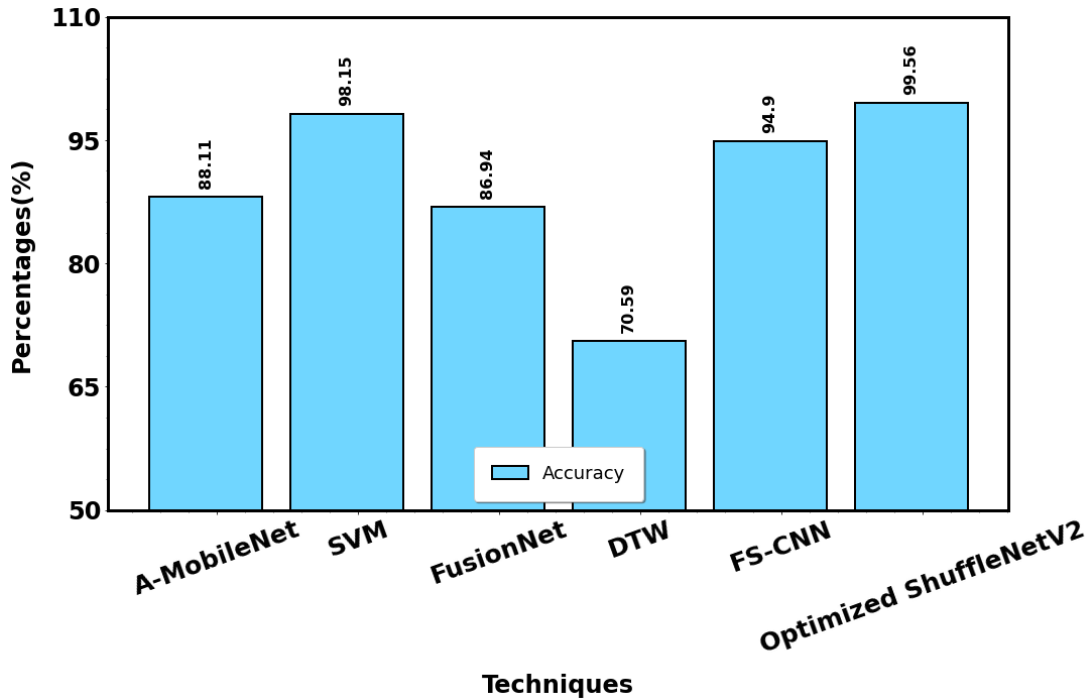


Figure 6: The Comparison Outcome of Proposed Method with Existing Methods

The proposed detection technique is accurately detecting the face part, so that the proposed novel deep learning technique classifies the face action and emotion correctly with higher amount of classification accuracy. A Figure 6 shows the comparison of proposed and existing methods on face action and emotion recognition.

E. Evaluation of Training and Testing:

After 100 epochs, the training accuracy and testing accuracy curves converge, yielding a respectable accuracy of 99.56%. Figure 7 shows the training and testing accuracy and loss for dataset 1.

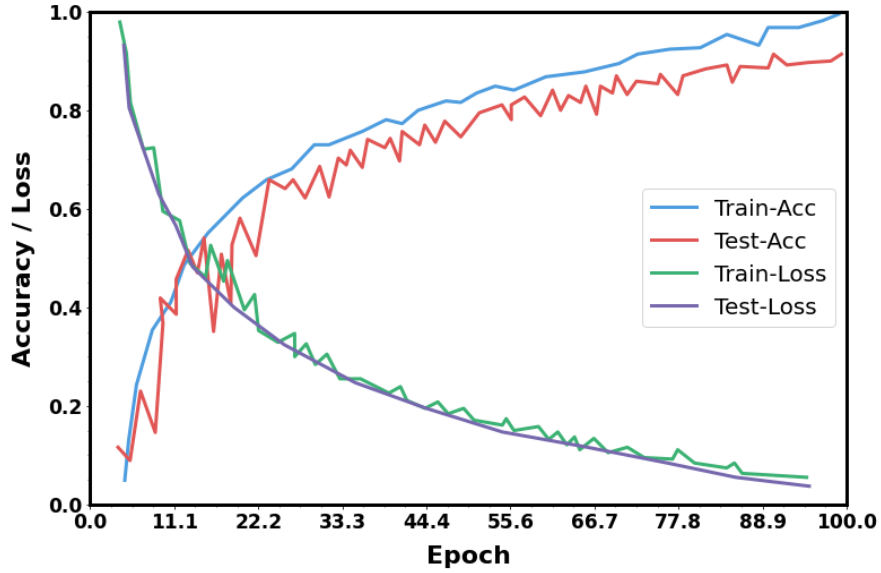


Figure 7: Training and Testing Accuracy and Loss for Dataset 1

The testing loss curve briefly oscillates up and down. Although the distinction between training and test loss is minimal and the curve does not increase across epochs, this might be acceptable. Figure 8 shows the training and testing accuracy and loss for dataset 2.

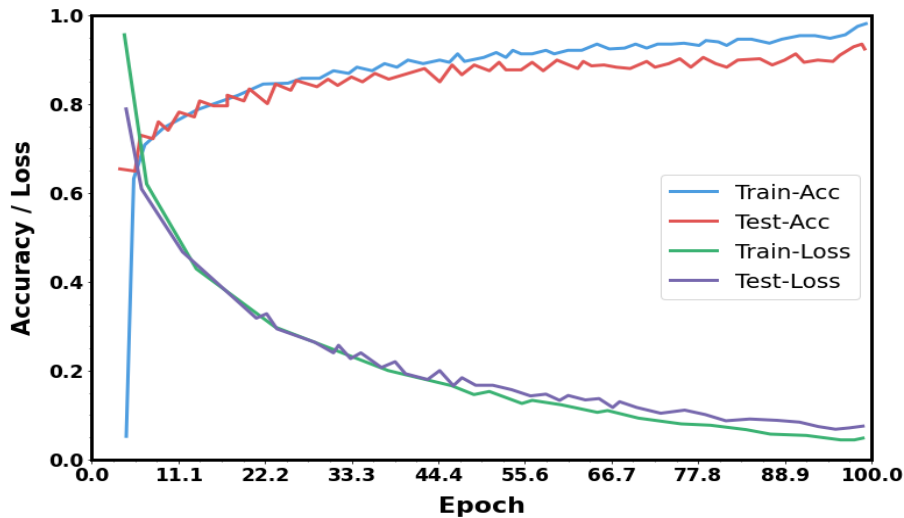


Figure 8: Training and Testing Accuracy and Loss for Dataset 2

The accuracy and loss during training are shown in Figure 7 and Figure 8. The Optimized ShuffleNetV2 is providing better accuracy and loss predictions. Compared to other techniques, our method is given better performance in the training and testing process for the classification of thermal image expressions and actions.

F. Computation Time:

Another aspect that is contrasted is computation time. Deep learning techniques try to handle computation complexity. When compared to other existing strategies, as demonstrated in Table 6, using the suggested Optimized ShuffleNetV2 technique requires less computational time. With minimal computing effort, it provides improved classification accuracy. The computing time needed to run the state-of-the-art methods and the proposed model on the datasets is shown in Figure 9.

Table 6: Computation Time of Proposed and Existing Methods

References	Techniques	Accuracy
Nan et al. [16]	A-MobileNet	0.26
Boughida et al. [17]	SVM	0.29
Litvin et al. [18]	FusionNet	0.19
Nayak et al. [19]	DTW	0.21
Said & Barr [20]	FS-CNN	0.27
Proposed Method	Optimized ShuffleNetV2	0.12

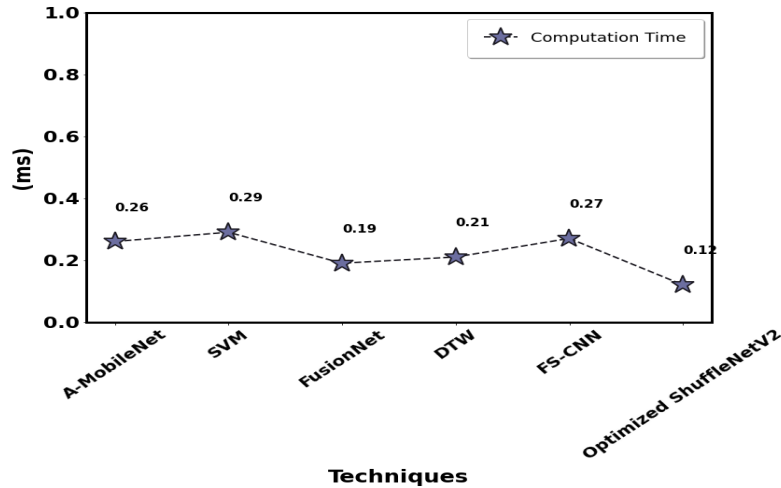


Figure 9: The Comparison of Proposed Method with Existing Method’s Computational Time Complexity

A categorization method's performance is evaluated using a confusion matrix. A confusion matrix is used to display and present the results of a method of classification. The computation times of the proposed approach and the existing methods are compared in Figure 9.

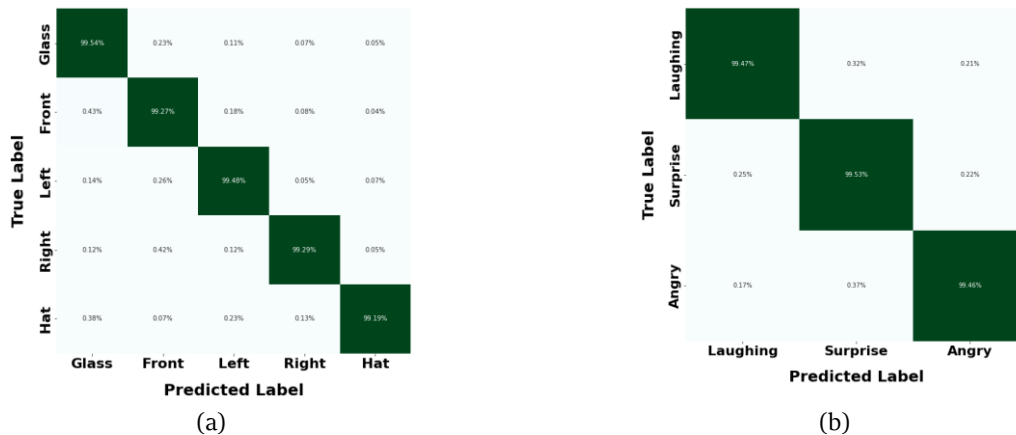


Figure 10: Confusion Matrix of Proposed Methodology on (a) Dataset 1, (b) Dataset 2.

A Figure 10 shows the confusion matrix of proposed methodology on dataset 1 and dataset 2.

V. CONCLUSION

Human action recognition involves the detection and classification of various actions performed by humans, typically from video data. These actions can include gestures, movements, and activities. We proposed a new deep learning approach to recognize the human actions and expressions in this paper. The relevant features are extracted through Improved Principal

Component Algorithm and after that, detected the human face using YOLOv8 approach. Finally, classify the human actions and expressions using ShuffleNetV2 method with the help of a novel optimization algorithm to improve the classification accuracy and it's performance with less computational time complexity. The proposed approach achieved 99.56% accuracy with less computation time. Compared to the existing method, the proposed achieved better. In future, we will recommend hybrid deep learning techniques to improve the performance of proposed study.

VI. REFERENCES

- [1] Prabhu, K., Kumar, S. S., Sivachitra, M., Dineshkumar, S., & Sathiyabama, P. (2022). Facial Expression Recognition Using Enhanced Convolution Neural Network with Attention Mechanism. *Computer Systems Science & Engineering*, 41(1).
- [2] Wang, K., Lian, Z., Sun, L., Liu, B., Tao, J., & Fan, Y. (2022, October). Emotional reaction analysis based on multi-label graph convolutional networks and dynamic facial expression recognition transformer. In *Proceedings of the 3rd International on Multimodal Sentiment Analysis Workshop and Challenge* (pp. 75-80).
- [3] Jeong, J. Y., Hong, Y. G., Kim, D., Jung, Y., & Jeong, J. W. (2022). Facial expression recognition based on multi-head cross attention network. *arXiv preprint arXiv:2203.13235*.
- [4] Savchenko, A. V., Savchenko, L. V., & Makarov, I. (2022). Classifying emotions and engagement in online learning based on a single facial expression recognition neural network. *IEEE Transactions on Affective Computing*, 13(4), 2132-2143.
- [5] Wang, Y., Sun, Y., Huang, Y., Liu, Z., Gao, S., Zhang, W., ... & Zhang, W. (2022). Ferv39k: A large-scale multi-scene dataset for facial expression recognition in videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 20922-20931).
- [6] Sadeghi, H., & Raie, A. A. (2022). HistNet: Histogram-based convolutional neural network with Chi-squared deep metric learning for facial expression recognition. *Information Sciences*, 608, 472-488.
- [7] Ruan, D., Mo, R., Yan, Y., Chen, S., Xue, J. H., & Wang, H. (2022). Adaptive deep disturbance-disentangled learning for facial expression recognition. *International Journal of Computer Vision*, 130(2), 455-477.
- [8] Moung, E. G., Wooi, C. C., Sufian, M. M., On, C. K., & Dargham, J. A. (2022). Ensemble-based face expression recognition approach for image sentiment analysis. *Int. J. Electr. Comput. Eng*, 12(3), 2588-2600.
- [9] Ma, T., Tian, W., & Xie, Y. (2022). Multi-level knowledge distillation for low-resolution object detection and facial expression recognition. *Knowledge-Based Systems*, 240, 108136.
- [10] Bodapati, J. D., Srilakshmi, U., & Veeranjanyulu, N. (2022). FERNet: a deep CNN architecture for facial expression recognition in the wild. *Journal of The institution of engineers (India): series B*, 103(2), 439-448.
- [11] Zhu, Q., Mao, Q., Jia, H., Noi, O. E. N., & Tu, J. (2022). Convolutional relation network for facial expression recognition in the wild with few-shot learning. *Expert Systems with Applications*, 189, 116046.
- [12] Nan, F., Jing, W., Tian, F., Zhang, J., Chao, K. M., Hong, Z., & Zheng, Q. (2022). Feature super-resolution based Facial Expression Recognition for multi-scale low-resolution images. *Knowledge-Based Systems*, 236, 107678.
- [13] Gan, C., Xiao, J., Wang, Z., Zhang, Z., & Zhu, Q. (2022). Facial expression recognition using densely connected convolutional neural network and hierarchical spatial attention. *Image and Vision Computing*, 117, 104342.
- [14] Liu, Y., Wang, W., Feng, C., Zhang, H., Chen, Z., & Zhan, Y. (2023). Expression snippet transformer for robust video-based facial expression recognition. *Pattern Recognition*, 138, 109368.
- [15] Zou, W., Zhang, D., & Lee, D. J. (2022). A new multi-feature fusion based convolutional neural network for facial expression recognition. *Applied Intelligence*, 52(3), 2918-2929.
- [16] Nan, Y., Ju, J., Hua, Q., Zhang, H., & Wang, B. (2022). A-MobileNet: An approach of facial expression recognition. *Alexandria Engineering Journal*, 61(6), 4435-4444.
- [17] Boughida, A., Kouahla, M. N., & Lafifi, Y. (2022). A novel approach for facial expression recognition based on Gabor filters and genetic algorithm. *Evolving Systems*, 13(2), 331-345.
- [18] Litvin, A., Nasrollahi, K., Escalera, S., Ozcinar, C., Moeslund, T. B., & Anbarjafari, G. (2019). A novel deep network architecture for reconstructing RGB facial images from thermal for face recognition. *Multimedia Tools and Applications*, 78, 25259-25271.
- [19] Nayak, S., Nagesh, B., Routray, A., & Sarma, M. (2021). A Human-Computer Interaction framework for emotion recognition through time-series thermal video sequences. *Computers & Electrical Engineering*, 93, 107280.
- [20] Said, Y., & Barr, M. (2021). Human emotion recognition based on facial expressions via deep learning on high-resolution images. *Multimedia Tools and Applications*, 80(16), 25241-25253.
- [21] Zhu, Y., Zhu, C., & Li, X. (2018). Improved principal component analysis and linear regression classification for face recognition. *Signal Processing*, 145, 175-182.
- [22] Talaat, F. M., & ZainEldin, H. (2023). An improved fire detection approach based on YOLO-v8 for smart cities. *Neural Computing and Applications*, 1-16.
- [23] Yang, R., Lu, X., Huang, J., Zhou, J., Jiao, J., Liu, Y., ... & Gu, P. (2021). A multi-source data fusion decision-making method for disease and pest detection of grape foliage based on ShuffleNet V2. *Remote Sensing*, 13(24), 5102.
- [24] Zhang, J., Zhang, G., Kong, M., & Zhang, T. (2023). Adaptive infinite impulse response system identification using an enhanced golden jackal optimization. *The Journal of Supercomputing*, 1-26.