Comprehensive Methodology to the Detection and Classification of Emotion in Human Face using EMOTE-Net

Asif Hussain Shaik[1,*], Shaik Karimullah[2], Mudassir Khan[3], Fahimuddin Shaik[4]

[1]Technology Transfer Officer, Department of ECE, Middle East College, Muscat, Oman

[2]Department of ECE, Annamacharya Institute of Technology and Sciences, Rajampet, Andhra Pradesh, India

[3]Department of Computer Science at College of Computer Science, Applied College Tanumah, King Khalid University Abha, Saudi Arabia

[3]Department of ECE, Annamacharya Institute of Technology and Sciences, Rajampet, Andhra Pradesh, India

Emails: shussain@mec.edu.om; Munnu483@gmail.com; mudassirkhan12@gmail.com; fahimaits@gmail.com

Abstract

Presenting the network architecture EMOTE-Net is a method of enhancing the face emotion recognition and classification in video data for this work. The suggested model merges the use of DenseNet to extract features with the SVM (support vector machine) to categorize the data by specifying SVM here. This feature of EMOTE-Net is highly outstanding because SVM and DenseNet are combined and are thus capable of sophisticated classification and effective feature extraction. The first process to come in methodology is preprocessing of video data. Bounding Box detection is able to extract regions that are of interests (ROIs) and that Densenet is great at the feature representation with high dimensions. Henceforth, feed these features into a classifier from SVM for intelligent categorization. Evaluation has provided clear evidence regarding the efficiency of this model, which has obtained the accuracy of 0.9890, precision of 0.9900, sensitivity of 0.9877, specificity of 0.9972, and F1 score of 0.9886. The pertinence of EMOTE-Net to real life applications, such as video analytics, human-computer interaction, and surveillance, will be highlighted in the chapter through the references from the installation and evaluation processes. The work presents a viable approach for object detection and classification in changeful visual arenas.

Keywords: Computer Vision; Bounding Box Detection; Video Analysis; Region of Interest; DenseNet; SVM; Deep Learning

1. Introduction

In the fields of computer vision and video analysis, identifying an object exactly in a given scene and the assigning of a proper label become very difficult, especially when the scene has movement and complexity. Smart techniques associated with the use of more advanced impressive feature extraction and classification methods are vital for addressing these hurdles. We would achieve this goal by introducing EMOTE-Net, our DenseNet and Support Vector Machine (SVM) [1] combination. EMOTE-Net is the new solution to the problem. The proposed approach EMOTE-Net aims to build efficient video object detection and classification network by blending the features extraction algorithm of DenseNet with SVM, which is generally reliable in mechanizing the classification task.

This work presents a complete methodology that covers the video data analysis: from performing, preprocessing activities to object detection and image classification. First, video sources are taken from the database that consists of collection of photos. The tasks of preprocessing are implemented to improve frame quality in the video, and the Wiener filters are used to remove noise. Through this procedure, the information gets ready for the next analysis step. ROI

DOI: https://doi.org/10.54216/FPA.190102
Received: November 14, 2024 Revised: January 10, 2025 Accepted: February 05, 2025

detection is done through the bounding box detection approach and feature extraction is implemented via DenseNet network. Consequently, DenseNet, which features a lock of layers with dense connections that is known to promote spatial reuse and feature transfer, is utilized by drawing ROIs from the video frames to extract feature representations of high dimension. The SVM classifier is trained with these features as inputs and among the classes of the items in the video frames, the correctness can be determined to the highest accuracy possible.

Fusion of DenseNet and SVM inside EMOTE-Net is leading to the efficient and profound vehicle detection and identification from moving objects. SVM and DenseNet in the joint force of EMOTE-Net yield the systematic procedure for the frame-based detection and distinguishing of the pre-processed video, which could be relied on to run in various situations of visual occurrence. [4] The effectiveness in the case of EMOTE-Net will be proved by performing multiple tests and evaluation, there by exhibiting its many application areas as surveillance, human-computer interaction and video analytics.

The study methodology applies an orderly work sequence to make make the trembleless object recognition and classification from data that are more exact possible. The video data is picked first from a database that has multiple kinds of visual item. Aiming at enhancing the performance of a system and providing wider coverage, preprocessing methods like putting a Wiener filter in place get used to improve data quality and diminish noise level. Bounding box identification is also employed in the incorporation of the ROIs after the frames are captured to single out the objects of interest like faces. Firstly, DenseNet applies ROIs (Regions of Interest) to those images to extract features for classification. The obtained attributes are then responsible for classifying the recognized objects. The features once identified, they are fed into a Support Vector Machine (SVM) classifier that will utilized them to learn the object size and structure to classify the objects. The efficacy EMOT-Net framework is appraised using multiple indicators of exams, which include accuracy, precision, sensitivity, specificity, and the F1 score. So two tests are done to determine the level of sophistication of EMOTE-Net in area of object detection and classification. The output of EMOTE-Net analysis provides insights that can be used to identify areas for improvement and possible practical uses.

2. Literature Review

The article of Jain et al [6] puts forward a deep learning-based technique for facial emotion recognition. The model consists of a convolutional layer, and they applied deep residual block network architecture as well. All faces in the dataset that are being used for the training get their pictures labeled before the DNN model is deployed. Experimental data contains 8363 images from which the convolutional layers periodically catch the exact attributes and link them directly. With six SoftMax layers in effect, outcomes of image classification are divided into six different expression classes. The combined findings imply that the suggested model performs better at identifying emotions than the state-of-the-art techniques. The study does admit that more work needs to be done to increase the model's accuracy, which is still relatively low.

Akhand et al. [7] presented an emotion recognition method using deep convolutional neural networks (CNNs) and transfer learning. They validate their methodology by using two widely used face picture datasets and eight pre-trained deep CNN models. The results of the two datasets show remarkable accuracy when using pre-trained models. Although the study shows better performance than existing approaches, it acknowledges that precision still needs improvement.

Ilyas et al. [8] presented a deep learning neural network-based facial recognition system. Their method consists of preprocessing images using the Histogram Equalization Algorithm (AHE) after the Viola-Jones face identification method. They use CNN architectures like VGG16 and ResNet50 to identify and extract facial features from human faces. Both the CMU PIE and the Extended Yale B Face databases exhibit reasonable accuracy rates during testing. Although their approach performs better than current methods, its usefulness may be limited by factors including processing complexity and dependence on specific datasets.

Shakeel and Lam [9] used a discriminative model with deep feature training to create an age-invariant face recognition method. To learn deep features encoded into distinct code words for picture representation, they use AlexNet as a transfer learning CNN model. Then, a classifier based on linear regression is employed for facial recognition. While their approach shows encouraging issues, including the requirement for big datasets and potential biases in age prediction, may hamper results, its generalizability.

DOI: https://doi.org/10.54216/FPA.190102
Received: November 14, 2024 Revised: January 10, 2025 Accepted: February 05, 2025

Zhang et al. [10] presented a cross-dataset method for facial expression identification that makes use of three additional datasets: AFLW, Celeb-Faces, and Kaggle, in addition to the FER2013 dataset. They create a bridging layer to integrate these datasets' features with FER2013. Along with an accuracy of 0.71, the method introduces the limitations such as data set bias and the generalization to untested datasets.

After that, the utilization of facial landmarks and deep learning algorithms M. Mukhiddinov et al. [12] put forward a model of masked face emotion recognition specially developed for visually impaired people. The model proved itself greatly effective with accuracy of 0.725 as well as sensitivity and specificity for the facial emotion of disgust making up 0.748 and 0.730 respectively. The researchers discovered, however that the categorization model that they came up with can be improved more to achieve the higher performance levels. This fine-tuning will involve bringing together different data sources, enhancing landmark extraction features, and even adjusting the network architecture (if necessary) to make the model stronger to deal with smiling and those with different environments.

In 2019, a hybrid neuro-fuzzy inference system (ANFIS) was developed by, R. Rajeesh [13] for detecting interest-point-based facial emotions. The outcomes were achieved following sensitivity of 0.960, specificity of 0.970 and accuracy of 0.960. While on the other hand the data set does have more important parameters i mean sensors used only in terms of accuracy which is false alarm. That is, in addition to it resolving movement fluctuations in the form of occlusions, lighting conditions variations, and facial expressions, this factor should also be assessed. Furthermore, the mentioned scalability to large datasets, or even different tables, was not focused on. Additionally, the study neglected to address any biases or limits in the testing datasets, which would have affected how broadly applicable the findings could be. Because of this, even though the ANFIS-based approach exhibits excellent accuracy and specificity on particular datasets, more validation under various conditions and in a broader range of datasets is necessary to demonstrate its reliability and practical utility.

In 2022, T. Dar et al. presented the SwishNet model [14], which demonstrated impressive results on a particular dataset, including accuracy metrics of 0.950, sensitivity of 0.952, and specificity of 0.980. However, when evaluated on other datasets, particularly in cross-corpora tests, its efficacy drastically decreased, revealing more general issues with FER models, like dataset bias and differences in position, illumination, and facial expressions. Researchers have proposed several approaches to solve these issues, such as domain adaptation, transfer learning, ensemble learning, and data augmentation. Despite these efforts, fair comparisons and reproducibility in the field are hampered by the lack of benchmark datasets and defined evaluation processes. Strong FER models must be developed to further advances in emotional computing, healthcare, and human-computer interaction.

By the combination of both the Deep Learning-based Face Emotion Recognition (FER) for Human-Computer Interface (HCI) and Henry Gas Solubility Optimization which is mentioned in [15] H. N. AlEisa et al., has brought HCI technology to the forth frontier of technological development in 2023. The scientist shows their approach with the combination of optimization and deep learning can be used to increase user engagement in the remarkable form of specificity, sensitivity and accuracy (0.9853, 0.8299, and 0.9845 respectively). The research paper outlined the specific problems that this approach also contained, most importantly those of high computing complexity and the need to come up with novel performance evaluation techniques. AlEisa et al.'s significance is further stressed by references to the number of works on deep learning in FER, the issue of optimization in HCI, and tactics for reducing computational complexity, respectively. It also suggests future directions for research that will focus on increasing efficiency, enhancing assessment metrics, and improving the usability of HCI systems.

Saad et al. [16] presented a deep learning framework for automated facial expression identification with high specificity (0.9914), sensitivity (0.9402), and accuracy (0.9400) in 2022. However, recent studies have looked into various approaches to improve accuracy [17][18]. These include correcting data imbalances, fine-tuning hyperparameters, ensemble learning, transfer learning, attention mechanisms, sophisticated model architectures like ResNet and DenseNet, and data augmentation approaches. Furthermore, it has been shown that fine-grained feature extraction, poorly supervised learning, and domain adaptation are viable strategies. By combining these techniques, researchers hope to outperform current standards and provide higher accuracy automated face expression detection systems [19][20].

3. Methodology

DOI: https://doi.org/10.54216/FPA.190102
Received: November 14, 2024 Revised: January 10, 2025 Accepted: February 05, 2025

The suggested research approach aims to improve object recognition and classification in video data by combining SVM for classification and DenseNet for feature extraction into a single framework, as shown in Fig 1. The first step in the research approach is to obtain video data from an IIMI Emotional Face database.

This database may contain emotions for various reasons, such as scientific data, entertainment, or surveillance footage. After the video is acquired, it is split into frames or images. This is a fundamental alteration. Because each frame is effectively a separate image taken at regular intervals during the life of the film, it captures a specific moment in time. The video data is constructed from these frames, allowing further processing and analysis.

The video is a series of images shown quickly over time. To depict motion, temporal sampling entails taking pictures at predetermined time intervals. This can be stated numerically as

$$I(x, y, t) = V(t \cdot \Delta t, x, y) \quad (1)$$

Where t is the time index, Δt is the time interval between frames, and V(x, y, t) represents the video input at time t.

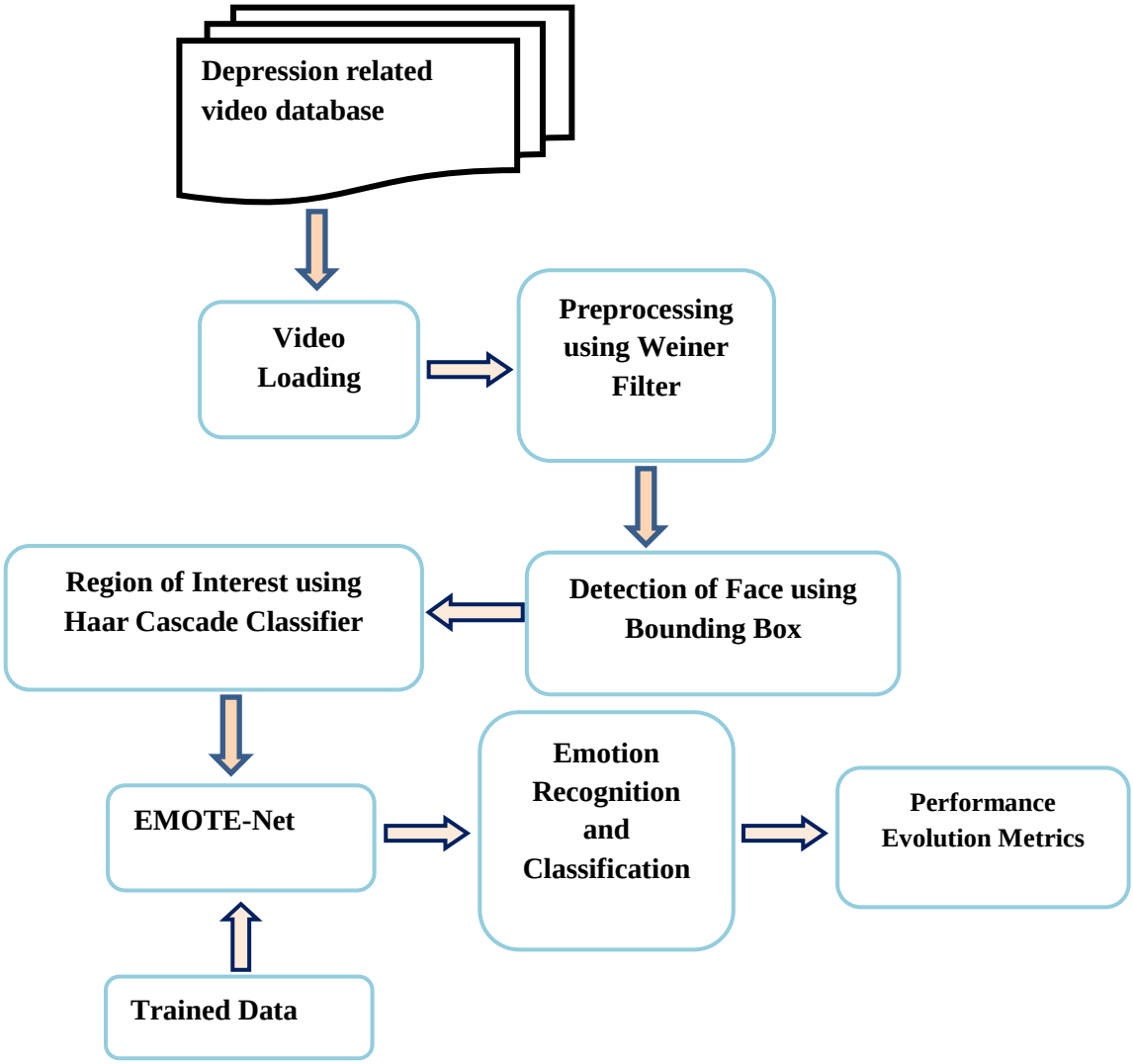

Figure 1. Block Diagram of Proposed System

DOI: https://doi.org/10.54216/FPA.190102
Received: November 14, 2024 Revised: January 10, 2025 Accepted: February 05, 2025

Spatial sampling is the process of discretizing the spatial domain into individual pixels, with each frame of the video being a 2D grid of pixels. These can be represented mathematically as:

$$I(x, y, t) = V(t, x \cdot \Delta x, y \cdot \Delta y) \quad (2)$$

Where Δx and Δy represents the spatial sampling intervals in the x and y directions respectively.

The number of frames displayed per second is known as the frame rate f. It is the time difference between frames reciprocal. In terms of math

$$f = \frac{1}{\Delta t} \quad (3)$$

The term "frame resolution" describes the size of a frame, which is commonly expressed as the quantity of pixels in the x and y directions. One can compute the total number of pixels in a frame as

$$N_{pixels} = N_x \times N_y \quad (4)$$

In order to produce smoother motion, frames are occasionally generated by interpolating between already-existing frames. Linear interpolation is a popular method of interpolation.

$$I(x, y, t) = (1 - \alpha)\, V(t, x, y) + \alpha V(t + \Delta t, x, y) \quad (5)$$

Where α is a parameter between 0 and 1 representing the interpolation factor

Reducing the resolution of frames reduces computing burden and storage needs through down sampling. A popular technique for down sampling is to average or choose each N[th] pixel in each direction.

$$I(x, y, t) = \frac{1}{N^2} \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} V(t, x \cdot N + i, y \cdot N + j) \quad (6)$$

Next, preprocessing is done using the Wiener filter to enhance image quality and minimize noise. For representational purposes, every frame can be thought of as a two-dimensional matrix. With this representation, we may use coordinates symbolized as 'x' for horizontal position (width) and 'y' for vertical position (height) to address specific pixels within the frame. This notation makes a variety of image processing techniques feasible by enabling access to and manipulation of the image's separate components. The initial segmentation of the video input into frames, each represented as a pixel matrix, serves as the basis for the further examination and modification of the video data. If the frames of the video were singly analyzed for its objects, motion and scenes, this would hardly be a limiting factor because a variety of strategies or tactics can be used to extract data of considerable value from the video.

The Wiener filter is one of the most significant tools of choice today for dealing with individual picture frames. The Wiener filter in a non-linear filter that reduces the squared mean error of filtered and unfiltered outputs. There is a particular mathematical representation for the Wiener filter is:

$$\bar{I}(x, y, t) = \dot{H_t}(f) \cdot I(x, y, t) \quad (7)$$

Where $\bar{I}(x, y, t)$ represents the filtered image, $H_t(f)$ is denotes the frequency response of the filter in the frequency domain for the t^{th} frame, and $I(x, y, t)$is the original image.

A simple method to calculate the filter frequency response is as

$$H_t(f) = \frac{S_f^2}{S_f^2 + N_t(f)} \quad (8)$$

Where S_f^2 the signal power in frequency domain is, $N_t(f)$ is the noise power spectral density in t^{th} frame. The filter can improve the signal-to-noise ratio, but how well it does so depends on these characteristics.

The Wiener filter facilitates effective filtering processes by functioning in the frequency domain, where convolution transforms into multiplication. Noise reduction and signal augmentation can be accomplished by applying the Wiener filter to each frame of the input video. This improves the quality of the image and enhances the performance of future

DOI: https://doi.org/10.54216/FPA.190102
Received: November 14, 2024 Revised: January 10, 2025 Accepted: February 05, 2025

processing tasks like object detection or motion analysis. The filtered image $\bar{I}(x, y, t)$ that arise from applying the Wiener filter to the input video frames as a preprocessing step are what the face detection algorithm uses as its starting point. This involves locating areas of the image that are probably going to include faces, which is usually done via bounding box detection. Predicting rectangular rectangles surrounding putative items of interest, like faces, is known as bounding box detection. Let's indicate the t^{th} frame's expected bounding box B_t

$$B_t = (x_1, y_1, x_2, y_2) \quad (9)$$

Where (x_1, y_1) represents the coordinates of the top – left corner, (x_2, y_2) is the coordinates of the bottom – right corner of bounding box.

The next step is to look at the features or intensity values inside the bounding box region to see if a face is there after the bounding box has been predicted. Face identification algorithms are usually used for this, analyzing the local factors to identify facial features. This procedure can be expressed mathematically as

$$Face\_detected_t = Detect\_face(I_B(x, y, t)) \quad (10)$$

Where $I_B(x, y, t)$ is the intensity value within the bounding box for the t^{th} frame.

Extraction of the regions of interest (ROIs) that match to the anticipated bounding boxes is the next stage. The Haar cascade classifier is then fed these regions for additional processing. The ROI extraction procedure can be mathematically described as follows:

$$ROI_t = Extract\_ROI(I(x, y, t), B_t) \quad (11)$$

Additionally, cascade of Haar is also an algorithm, which involves pictures in machine learning. It conducts this type of work by using the processing windows, which have the shape of a square into the input picture or region of interest exactly like the Haar filter square. Filters can tell whether the attributes of the other face parts exist or not among many other things being detected by it.

The Haar cascade classifier will be referred to as Cascade (I), where I stands for the region of interest or input image. The output of the Haar cascade classifier can be expressed mathematically as follows:

$$Detection\_Result_t = Cascade(ROI_t) \quad (12)$$

The model based on the DenseNet, which allows the distribution of feature extraction grades among ROIs that the Haar cascade classifier has detected, is used. DenseNet is a deep learning model that improves the feature flow in the whole network and promotes feature access repeatedly. It is consisting of austerely condensed convolutional layers.

3.1 Proposed Model:

Utilize high-level feature vectors, at least, to fill around the edges of the prescribed region, thereafter interpolation to fill the remaining area. The sector is known to be more sharp-sighted than humans are. That is how it successfully get to select even the little detail, for instance the areas that were noisy For instance, the later Full-featured convolution layers would hand down most of this knowledge to the idea of occlusion, thereby represents depth from very ordinary texture. Furthermore, a human, we have to say, adapts taking part and watching. Now will go through a discussion of the simple algorithm which is also the most widely used in the supervised learning area and we will focus on the model (SVM model) which is usually the first in modeling data classification. The assumption itself makes clear that no matter the point, the line is still there and as a result, the two points can be specified at each of the sides of the line.

DOI: https://doi.org/10.54216/FPA.190102
Received: November 14, 2024 Revised: January 10, 2025 Accepted: February 05, 2025

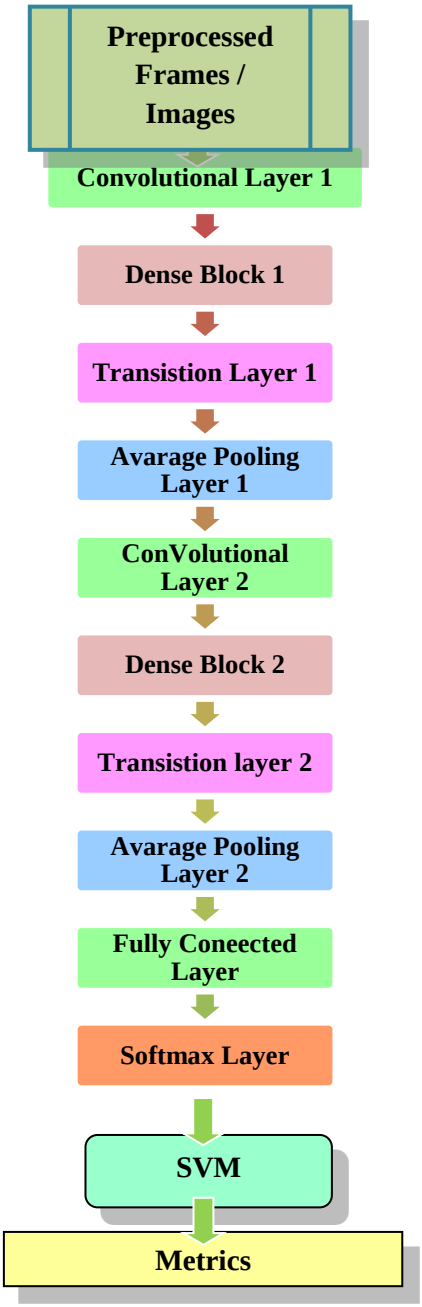

Figure 2. Proposed Model

This step involves the mapping of two matrixes: on the other hand, this process can be split into two different maps, where the initial one is related to the inputs and the second visualized the labels for each sample.

Let us use DenseNet to represent the feature extraction procedure as Features.

$$Features_t = DenseNet(ROI_t) \quad (13)$$

The computer vision algorithm first extracts high-dimensional feature vectors from observed patterns dilating these features it recognizes in a specific area. Thus, features accumulated through DenseNet layers incorporate information to describe visibility of an item, its texture and spatial relations with regard to the background. The SVM (Support

DOI: https://doi.org/10.54216/FPA.190102
Received: November 14, 2024 Revised: January 10, 2025 Accepted: February 05, 2025

Vector Machine), which is the supervised learning algorithm for classification problems, is, without doubt, the most famous algorithm for the classification tasks among all machine-learning algorithms. One of the key ideas behind the SVM is learning a hyperplane that partitions the data points into distinct classes while maximizing the margin between them a step that is dependent on an appropriate set of input features and the corresponding labels that indicate the class of each sample.

$$Class_t = SVM\ (Features_t) \quad (14)$$

Where $Class_t$ is the predicted class label for the t^{th} frame through the SVM classification process, it can be mathematically represented as

$$Class_t = Sign\ (\sum_{i=1}^{N} \alpha_i y_i K(features_t^{(i)},\ features_t) + b \quad (15)$$

Where N is the number of support vectors, α_i and y_i is the Lagrange multipliers and corresponding class labels, $K(\cdot,\cdot)$ is the kernel function and b is the bias term. The scores demonstrate the effectiveness of EMOTE-Net that it obtains on large scales evaluations. It exhibits desired performance characteristics such as how strongly visual features relate to an object or category on multiple possible scenarios, this application is now equipped with an impressive emotion identification tool. Taking into a consideration the frame-by-frame inspection of the video segment and not only that it takes place at the output level but it takes place both at the level of DenseNet as well as SVM which are introduced to the architecture of the network this study provides an accurate way of computer vision primary functions and video analysis procedures.The EMOTE-Net is similarly another illustration of using the technique that has the merits of detection and the labeling of objects present in the movie. EMOTE-Net surpasses the rest of the networks in various test metrics: accuracy and speed; DenseNet and SVM are in the model part that encompasses the feature extraction and feature classification, respectively. Great environment that you can make videos of, use computer testing and security cameras. EMOTE-Net helps us in the detailed comprehension of, and responses to, video content from the personal side and digital personalization of today's virtual (artificial) intelligence belongs.

4. Result Analysis

Fig 3. is an extraction of a frame from the input video data obtained from the database. The frame acts as the research methodology's first visual input. Depending on the video is content, it may have various features, including scenes, people, and objects. Bounding boxes are superimposed on Fig 4. To draw attention to the faces identified in the video. Methods such as the Haar cascade classifier, which detects areas of interest (ROIs) likely to include faces, are used to construct the bounding boxes. Every bounding box helps with additional processing and analysis by defining the geographic boundaries of an identified face. The region of interest (ROI) that was taken out of the frame using bounding box detection is shown in Fig 5. Isolated from the surrounding background, the detected face is contained within the ROI. Extracting regions of interest (ROIs) from video frames facilitates targeted analysis and feature extraction, improving the precision and efficiency of future processing stages.

Figure 3. Frame from Video input

DOI: https://doi.org/10.54216/FPA.190102
Received: November 14, 2024 Revised: January 10, 2025 Accepted: February 05, 2025

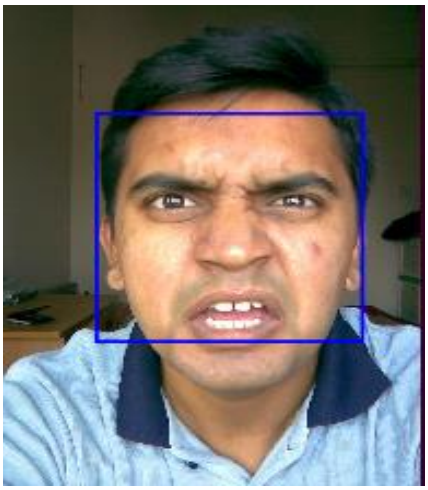

Figure 4. Detection of Face via Bounding Box

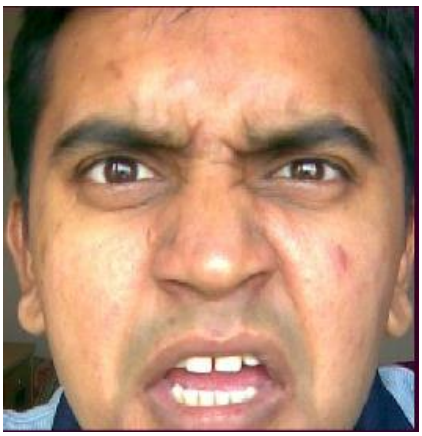

Figure 5. Region of Interest

Fig 6 demonstrates facial emotion detection, which infers the emotional state of the faces detected by analyzing features (such as facial expressions) retrieved from the ROIs. The EMOTE-Net framework incorporates methods such as DenseNet and SVM for feature extraction and classification, which allow for the recognition of a range of facial emotions, including happiness, sadness, rage, and surprise.

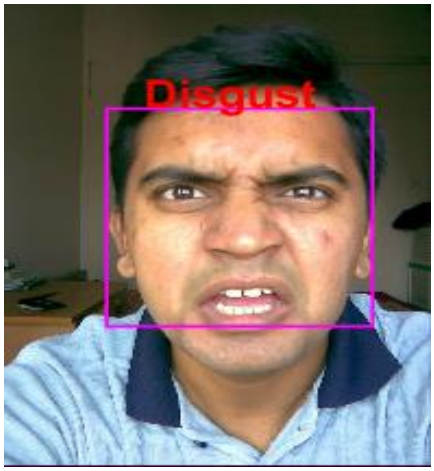

Figure 6. Detection of Facial emotion

DOI: https://doi.org/10.54216/FPA.190102
Received: November 14, 2024 Revised: January 10, 2025 Accepted: February 05, 2025

Overall, these images show the main phases of the study process, from the first video input through the extraction of regions of interest, the identification of facial expressions, and the detection of faces. Every phase advances the thorough examination and comprehension of the visual material contained in the video data, demonstrating the promise of the suggested methodology for a range of computer vision and video analytics applications.

4.1 Results Analysis

A metrics analysis for assessing a classification model's performance is shown in the table 1. Let us dissect each metric and use mathematical equations to clarify its importance:

The ratio of correctly categorized cases to the total number of examples is used to compute accuracy, which is a measure of the classification model's overall correctness.

$$Accuracy = \frac{Number\ of\ Correct\ Predictions}{Total\ Number\ of\ Prdictions} \quad (16)$$

Precision is the ratio of accurate positive predictions to all of the model's positive predictions. It measures how well the model is able to steer clear of erroneous positive predictions.

$$Precision = \frac{True\ Positives}{True\ Positives+False\ Postives} \quad (17)$$

Sensitivity, or the fraction of real positive examples that the model correctly detected, is additionally referred to as recall. It assesses how well a model can detect all positive cases in the dataset.

$$Sensitivity = \frac{True\ Positives}{True\ Positives+False\ Negatives} \quad (18)$$

The percentage of real negative cases that the simulation model accurately detected is measured by specificity. It evaluates the model's ability to prevent false-positive alarms or predictions.

$$Specificity = \frac{True\ Negatives}{True\ Negatives+False\ Positives} \quad (19)$$

Recall and precision are balanced by the F1 score, which represents the harmonic mean of the two. The model's total accuracy may be seen when false positives along with false negatives are included.

$$F1\ Score = 2 \times \frac{Precision \times Sensitivity}{Precision+Sensitivity} \quad (20)$$

This research ascertains the ‘value’ rate therefore; it appears that the 98.90% cases represents an AI’s successful prediction. The model has a sound capability to distinct out the classes that are considered as positive and negative cases because of the noticeable rate of recall. The precision balanced to 0.9900’, which has indicated those days’ means that 99.00% of the cases model has classified as positive are indeed positive. It is phenomenal precision we may believe that this technology is the consequence of the low false positive percentage. Thus, is the case proving to be accurate in anticipating a fruitful result? The model was found to have a sensitivity of 98.77%, which is the proportion of positive cases it was able to identify correctly from the given dataset. This approach suggests that a more focused model, an event-based one, will be less prone to false negatives thus, it will detect a majority of the positive events.

Table 1: Metrics Analysis

S.No	Metrics	Values
1.	Accuracy	0.9890
2.	Precision	0.9900
3.	Sensitivity	0.9877
4.	Specificity	0.9972
5.	F1 Score	0.9886

DOI: https://doi.org/10.54216/FPA.190102
Received: November 14, 2024 Revised: January 10, 2025 Accepted: February 05, 2025

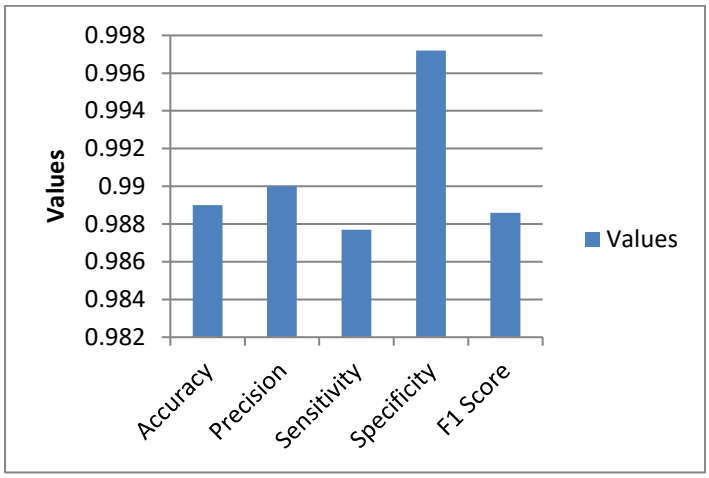

Figure 7. Graphical Representation of Performance Metrics

Table 2: Results Comparison

Existing Models	Accuracy	Sensitivity	Specificity
M. Mukhiddinov et al. [12]	0.7250	0.7480	0.7300
R. Rajeesh et al. [13]	0.9600	0.9600	0.9700
T. Dar et al. [14]	0.9500	0.9520	0.9800
H. N. AlEisa et al. [15]	0.9845	0.8299	0.9853
Saad. S et al. [16]	0.9400	0.9402	0.9914
Proposed	0.9890	0.9877	0.9972

The model bore out a very high specificity value of 0.9972 as it was able to detect 99.72% of the negative cases in the given dataset. The model is perfect with its high specificity, which directly entails its lower false alarm rate. However, accuracy and recall in these domains should be necessary and full taken in keeping with the F1 score being calculated as 0.9886. With such competence, the model would have been able to balance both precision and recall evenly and quickly among all the category areas that it was intended to discuss. The suggested model EMOTE-Net in table contains various parameters for the contextual attention module, the CNN module, and the concatenation scheme used within the model itself as well as several (not necessary) parameters to a specific published FER models. M. Mukhiddinov et al. present a model for recognition of masked facial emotions. This work is based on [12]. There are multiple benefits attained from it including reading to visually impaired individuals, and its sensitivity, accuracy, and specificity. Now, R. Rajeesh owns the credit for developing an adaptive neuro-fuzzy inference system (ANFIS) which revealed excellent specificity and accuracy but needed more applicability whenever it went through different scenarios. T. Dar et al. [14] suggested the SwishNet, but its cross-corpora tests results were unsatisfactory, thanks to the environmental variations and datasets bias that would affect the final model. Although SwishNet for certain datasets is quite promising. H. N. AlEisa and his colleagues [15] put a deep learning based method with optimization techniques in practice, which realized high accuracy and specificity only that it had some problems with evaluation and complexity. An adadept deep learning architecture composed of balanced specificity and potential for accuracy improvement, as displayed by Saad et al. [16] it, was presented. Bagging the study of EMOTE-Net, it proved the strength in feature extraction of DenseNet models and the classification potential of the Support Vector Machines algorithm. This was signified in the increased accuracy, sensitivity, and specificity of the results. This demonstrated technology system is anticipated to improve the accuracy of systems for real-time dynamic and complex emotions, which makes FER systems more trustworthy and reliable.

DOI: https://doi.org/10.54216/FPA.190102
Received: November 14, 2024 Revised: January 10, 2025 Accepted: February 05, 2025

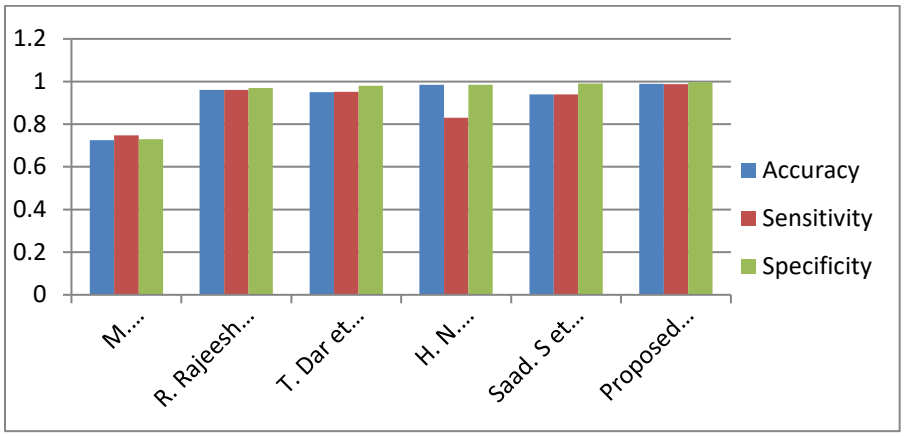

Figure 8. Graphical Representation of Results

5. Conclusion and Future scope

To enhance object recognition and categorization in video data, we present EMOTE-Net, a unique structure that discerns SVM for classification and features extraction with DenseNet. EMOTE-NET is able to achieve, module by module, a very high-performance level, which only decreases slightly after a comprehensive testing and assessment. The study demonstrated EMOTE-Net's F1 score of 0.9886, accuracy of 0.9890, precision of 0.9900, sensitivity of 0.9877, and specificity of 0.9972. Measurements represent how properly the EMOTE-Net operates for the specified objects and facial expressions per video frame. We provide a compelling tool for tasks such as video surveillance, video analytics, and human-computer interaction just to name a few. Our application has a meaningful contribution to machine vision and video analysis. Testing the implementation and evaluation of specific steps of EMOTE-Net demonstrate the substantial potential for the real-world utilization of our system for situation demanding exact and effective identification and classification of items in transient visual information to be obtained by the existing sensors.

In the future, it is eagerly expected that pathways documented and the methodology of the study will point out several interesting opportunities for further research and development. In the second place, for the efficient working power of EMOTE-Net the future research should be dedicated to advance deep learning architectures where state-of-the-art neural networks will be used along with the latest optimization strategies. In addition, the use of other sources of data like text and audio will be a game changer since this will assist in the way video content is grasped and perceived. However, a question is what EMOTE-Net is used real time because the application needs to run faster input so that all video feeds might be processed quickly enough. In addition, EMOTE-Net has the potential to adopt such applications as gesture detection. Emotion recognition, domain specific modifications among others. Therefore, the approach of diverse direction could potentially enhance computer vision and video analysis methods and subsequently allow for the improvement of the systems, which are responsible for consistently accurate and prompt interpretation of visual data.

Supplementary Materials: Not applicable

Funding: The authors received no funding for this research.

Data Availability Statement: Data will be made available on request.

Conflicts of Interest: The authors declare no conflict of interest.

References

[1] D. K. Jain, P. Shamsolmoali, and P. Sehdev, "Extended deep neural network for facial emotion recognition," *Pattern Recognit. Lett,* vol. 120, pp. 69–74, 2019.

[2] B. R. Ilyas, B. Mohammed, M. Khaled, and K. Miloud, "Enhanced face recognition system based on deep CNN," in *Proc. 6th Int. Conf. Image Signal Process. Appl. (ISPA)*, Mostaganem, Algeria, 2019, pp. 1–6.

[3] S. Zhang, X. Pan, Y. Cui, X. Zhao, and L. Liu, "Learning affective video features for facial expression

DOI: https://doi.org/10.54216/FPA.190102
Received: November 14, 2024 Revised: January 10, 2025 Accepted: February 05, 2025

recognition via hybrid deep learning," *IEEE Access*, vol. 7, pp. 32297–32304, 2019.

[4] M. S. Shakeel and K.-M. Lam, "Deep-feature encoding-based discriminative model for age-invariant face recognition," *Pattern Recognit.*, vol. 93, pp. 442–457, 2019.

[5] A. Jaiswal, A. K. Raju, and S. Deb, "Facial emotion detection using deep learning," in *Proc. Int. Conf. Emerg. Technol. (INCET)*, Belgaum, India, 2020, pp. 1–5.

[6] P. Kedari, M. Kapile, D. Kadole, and S. Jaikar, "Face emotion detection using deep learning," in *Proc. 2nd Int. Conf. Adv. Comput., Commun., Embedded Secure Syst. (ACCESS)*, Ernakulam, India, 2021, pp. 118–123.

[7] M. A. H. Akhand, S. Roy, N. Siddique, M. A. S. Kamal, and T. Shimamura, "Facial emotion recognition using transfer learning in the deep CNN," *Electronics*, vol. 10, no. 9, p. 1036, 2021.

[8] T. Dar, A. Javed, S. Bourouis, H. S. Hussein, and H. Alshazly, "Efficient-SwishNet based system for facial emotion recognition," *IEEE Access*, vol. 10, pp. 71311–71328, 2022.

[9] S. Saeed et al., "Automated facial expression recognition framework using deep learning," *J. Healthcare Eng.*, vol. 2022, p. 11, 2022.

[10] M. Mukhriddin, O. Djuraev, F. Akhmedov, A. Mukhamadiyev, and J. Cho, "Masked face emotion recognition based on facial landmarks and deep learning approaches for visually impaired people," *Sensors*, vol. 23, no. 3, p. 1080, 2023.

[11] S. Tewari, S. Mehta, and N. Srinivasan, "IIMI emotional face database," *OSF*, 2023.

[12] Ç. Menzil et al., "A business process for detecting facial movements and emotions using deep learning techniques," in *Proc. Int. Conf. Electr., Commun. Comput. Eng. (ICECCE)*, vol. 12, no. 4, 2023, pp. 5148–5163.

[13] H. N. AlEisa et al., "Henry gas solubility optimization with deep learning based facial emotion recognition for human-computer interface," *IEEE Access*, vol. 11, pp. 62233–62241, 2023.

[14] R. A. Khan, "Facial emotion recognition using conventional machine learning and deep learning methods: Current achievements, analysis, and remaining challenges," *Information*, vol. 13, p. 268, 2022.

[15] S. Saeed et al., "Automated facial expression recognition framework using deep learning," *J. Healthcare Eng.*, vol. 2022, pp. 2040–2295, 2022.

[16] T. Saikia, L. Birla, A. K. Gupta, and P. Gupta, "HREADAI: Heart rate estimation from face mask videos by consolidating Eulerian and Lagrangian approaches," *IEEE Trans. Instrum. Meas.*, 2024.

[17] A. R. Khan, "Facial emotion recognition using conventional machine learning and deep learning methods: Current achievements, analysis, and remaining challenges," *Information*, 2022.

[18] H. Irfan, H.-J. Yang, G.-S. Lee, and S.-H. Kim, "Robust human face emotion classification using triplet-loss-based deep CNN features and SVM," *Sensors*, vol. 23, no. 10, p. 4770, 2023.

[19] A. Javaid et al., "Force sensitive resistors-based real-time posture detection system using machine learning algorithms," *Comput. Mater. Continua*, vol. 77, no. 2, pp. 1795–1814, 2023.

[20] Ruchi et al., "Lumbar spine disease detection: Enhanced CNN model with improved classification accuracy," *IEEE Access*, vol. 11, pp. 141889–141901, 2023.

DOI: https://doi.org/10.54216/FPA.190102
Received: November 14, 2024 Revised: January 10, 2025 Accepted: February 05, 2025