International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10
***Article Received:** 12 August 2023 **Revised:** 22 September 2023 **Accepted:** 10 October 2023*

A Diaspora of Humans to Technology: VEDA Net for Sentiments and their Technical Analysis

[1]Kirti Sharma, [2]Rainu Nandal, [3]Shailender Kumar
[1]CSE Department, University Institute of Engineering and Technology,
Maharshi Dayanand University Rohtak Haryana India
Email: krtbhardwaj1@gmail.com
[2]CSE Department, University Institute of Engineering and Technology,
Maharshi Dayanand University Rohtak Haryana India
Email: rainunandal11@gmail.com
[3]Department of Computer Science, Delhi Technological University, New Delhi
Email: shailenderkumar@dce.ac.in

Abstract: ***Background:*** Human sentiments are the representation of one's soul. Visual media has emerged as one of the most potent instruments for communicating thoughts and feelings in today's world. The area of visible emotion analysis is abstract due to the considerable amount of bias in the human cognitive process. Machines need to apprehend better and segment these for future AI advancements. A broad range of prior research has investigated only the emotion class identifier part of the whole process. In this work, we focus on proposing a better architecture to assess an emotion identifier and finding a better strategy to extract and process an input image for the architecture.

Objective: We investigate the subject of visual emotion detection and analysis using a connected Dense Blocked Network to propose an architecture VEDANet. We show that the proposed architecture performed extremely effectively across different datasets.

Method: Using CNN based pre-trained architectures, we would like to highlight the spatial hierarchies of visual features. Because the image's spatial regions communicate substantial feelings, we utilize a dense block-based model VEDANet that focuses on the image's relevant sentiment-rich regions for effective emotion extraction. This work makes a substantial addition by providing an in-depth investigation of the proposed architecture by carrying out extensive trials on popular benchmark datasets to assess accuracy gains over the comparable state-of-the-art. In terms of emotion detection, the outcomes of the study show that the proposed VED system outperforms the existing ones (accuracy). Further, we explore over the top optimization i.e. OTO layer to achieve higher efficiency.

Results: When compared to the recent past research works, the proposed model performs admirably and obtains accuracy of 87.30% on the AffectNet dataset, 92.76% on Google FEC, 95.23% on Yale Dataset, and 97.63% on FER2013 dataset. We successfully merged the model with a face detector to obtain 98.34 percent accuracy on Real-Time live frames, further encouraging real-time applications. In comparison to existing approaches, we achieve real-time performance with a minimum TAT (Turn-around-Time) trade-off by using an appropriate network size and fewer parameters.

Keyword: Sentiment Analysis, Emotion Detection, Facial Emotion Recognition, FER, Deep Learning Architectures.

I. INTRODUCTION

Many research disciplines, such as computer vision, psychology, medical science and political science, rely on the identification of emotional state of people. Emotions have an influence on how information is received, how people create attitudes, and how they make decisions.

Everyone believes that sentiments exist on a deeper level of the human psyche, whereas emotions at a particular moment arise on the basis of the subject's sentiment for a particular object or a person. Sentiments, generally, become a part of human character which stays with the human for a longer period of time, whereas emotions are generally expressed for a shorter period of time (usually momentarily) as a result of a particular action or event relating to a particular sentiment of a human, in-turn triggering an emotional inburst or outburst.

During social interactions and confrontations with others, the face serves as a display board on which emotions and intentions are expressed, and it is scrutinised by others. Cognitive human emotion recognition and machine intelligence (AI), often known as human-computer interaction, are two distinct subfields of the larger academic area of emotion detection. There are variety of ways to express human's emotional state, including facial landmarks, body language (non-verbal) [1] and speech (vocal or verbal) [2]. Visual information made up 55% of emotional information, audio information 38%, and verbal information 7%, according to Mehrabian's research from 1967 [3]. These findings, more or less, stands valid till now. Facial variations during communication are the earliest evidence of emotional state transmission, that's why this medium (visual) has piqued the interest of most studies.

Study of facial expressions and variations will provide a base for advanced robotic and operative solutions. Robotic medical procedures to advanced autonomous driver assistance solutions will all rely on human aspect-based information extraction-processed and categorized into logical units for technology to work upon Comparing and categorizing facial traits is a complex and intricate process. In 1978, Ekman and Freisen pioneered the study of facial expressions and developed the Facial Action Coding System (FACS). This system characterizes facial movements through Action Units (AU's). Action Unit is the inclusion of one or more facial muscles which are sub-divided into 46 Action Units [4].

There are already a number of conventional techniques available for the extraction of face features, including geometric features like Gabor wavelets [6], local binary patterns [5], and facial action units (FACs) [4]. However, classifying human emotions into categories and precisely determining the true emotion are highly challenging tasks for a machine. Compared to other modalities, researchers have found that Automatic Facial Emotion Detection is the most investigated domain [7], but the work is difficult because everyone expresses emotion differently. It is analyzed that researchers should not ignore a number of additional hurdles and obstacles that exist in this domain, such as variations in head posture, luminance, background, and occlusion. Face landmark models and augmentation techniques have made significant progress in addressing many challenges, but there is always room for further improvement.

The success and effectiveness of Deep Neural Networks (DNN) have garnered significant attention from both corporations and academia in recent times. Researchers are using deep architectures such as the deep convolutional neural network, recurrent neural network, and attention networks for automatic classification and feature extraction to understand human emotions. Researchers have made considerable attempts to construct deep network designs and even integrated blocks to further improve the accuracy like Squeeze-and Excitation Block. Hu et al. (2018) presented a squeeze and excitation network (SENet) on top of a deep residual network to use the channel-wise attention mechanism. The attention-module enhances the residual block by rearranging the channels such that the residual block may learn quality characteristics [8].

Liu et al. (2019) introduced a framework based on ELM with SVM approach that obtained the accuracy of 85% on CK++ dataset [9] and employed Local Binary Pattern (LBP), 2D-Gabor to extract feature discrimination. In order to identify the 7 basic emotions in real-time facial emotion classification, Deeb et al. (2022) presented the human emotion recognition framework combining HAAR and LDA approach with IBH-based Extreme Learning Machine (ELM) classifier. The algorithm's accuracy on the KDEF, CK+, and JAFFE (Japanese Female Facial Expression) public datasets was 91.32 %, 97.62 %, and 97.78 %, respectively [10]. However, appearance features do not work well with noisy facial images and are inconsistent with a variety of facial variations (such as the dimension of the face, the orientation of the head region, and the appearing face region etc.). By taking a cue from the most recent works in this area., we also proposed a framework which possibly offers a variety of social uses and can be used to build a lot of real time applications in variety of fields, including education, automatic vehicles, entertainment, security surveillance, and others.

Some notable pre-trained CNN models, such as as ResNet and VGGNet are generally utilized to address the issue of excessive workload on bigger data training. There hasn't been any forward-thinking work with high efficiency in the field of Visual Emotion Detection deploying a pre-trained model to generate outcomes that are as accurate as possible, such as actual human emotions or sentiments. In our framework, we considered the ResNet as baseline architecture for the task of visual feature extraction.

The concept of *transfer learning* in computer vision inspired the feature extraction process by training deep CNNs on huge data sets like ImageNet for (VSA). This eliminates the requirement to train a model from the ground up. Likewise, we applied transfer learning (TL) in our proposed work to solve this problem and prevent the overfitting issue by transferring information from one domain to another.

The key research objectives of the proposed work for the present paper are further illustrated below.

A. Research Objectives:

The major objective of the presented research work is to provide a uniform framework for visual emotion detection in facial images and live streams along with 7 distinct emotions: Happy, anger, disgust, neutral, fear, sad, and Surprised.

The following is a summary of the key contribution of the proposed work:

- Development of an efficient face detection and recognition model with the help of pre-trained architectures to extract the visual face descriptors (FD’s). The visual face descriptors (FD’s) along with Bounding Boxes and 68 Landmark points based on *modified Resnet* (31layers) architecture leads the model up to high recognition accuracy.
- Proposed VEDANet architecture (baseline: DenseNet-121) for Visual Emotion Detection and Analysis and further investigate the proposed model on benchmark datasets as well as on real-time live frames.
- Extended the proposed architecture and add OTO Layer (Over the Top Optimization Layer) in order to improvise the resultant confidence score and to accelerate the emotion recognition accuracy rate.
- Last stage of this work concluded with comparison of the proposed model’s (VEDANet) emotion recognition accuracy with existing SOTA and real time live frames, proving it’s competency to outperform.

B. Datasets Used:

Extended Yale Face Database: This collection is made up of 16,128 facial pics that were all captured using the same light source. It has 28 unique subjects for 576 different viewing settings, including nine poses for each of the 64 different lighting conditions. Each facial image's actual dimensions are 320 x 243 pixels [11].

Google FEC Database: 700,000 triplets of distinct face crop photos made up the Google Facial Expression Comparison (FEC) [12] database. In each triplet, the most comparable pair of facial expressions are annotated.

Figure 1. Samples from Google Facial Expression Comparison Dataset

Affectnet Dataset: AffectNet [13], is actually the largest publicly accessible repository of face emotions with over 440,000 wild facial images, every photo is meticulously labelled by hand into eight categories of facial expression categories (Neutral, Happy, Fear, Sad, Surprise, Disgust, Contempt and Anger). In the presented work, to generalize the model seven emotion classes with annotations used except contempt emotion.

Figure 2. Example images from the AffectNet dataset. Emotion classes are labeled as 0 (neutral), 1 (happy), 2 (sad), 3 (surprise), 4 (fear), 5 (disgust), 6 (anger), and 7 (contempt).

FER2013: The FER-2013 dataset, pioneered by Pierre Luc Carrier and Aaron Courville, was crafted specifically for the facial expression recognition challenge at the ICML 2013 workshop [14]. This robust collection was meticulously curated by harnessing the Google image search API to gather photographs linked to emotion-centric keywords.

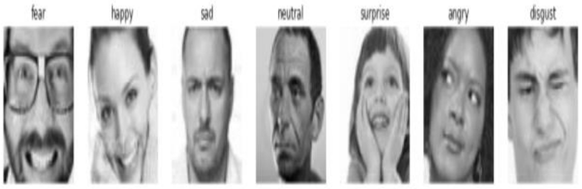


Figure 3. Representative images from the FER2013 Dataset, (total of 35,887 facial images broken down as: '8989 Happiness' + '6198 Neutral' + '6077 Sadness' + '5121 Fear' + '4953 Anger' + '4002 Surprise' + '547 Disgust')

Real-Time wild Dataset: We used a camera device (smartphones and webcam) to capture ongoing conversations and extracts ~350 frames from random instances. We exposed those frames to our proposed architecture to evaluate.

II. BACKGROUND AND RELATED WORK:

Emotion detection, as a field, has witnessed transformative advancements, predominantly with the integration of Convolutional Neural Networks (CNNs). The inherent design of CNNs, with their capability to process and categorize visual inputs, positions them at the forefront of this domain. A fundamental CNN is architected with a mosaic of neurons organized into layers, primarily encompassing convolutional, pooling, and fully connected ones. While the first two layers prioritize feature extraction and reduction to avert overfitting, the latter aids in classification, harnessing attributes from preceding layers. Notably, architectures like Inception V3 [15] have strategically embedded a Global Mean Pooling operation, streamlining the feature map into a scalar vector. Moreover, the advent of techniques such as residual modules [16] and depth-wise-separable convolutions [17] signify steps towards parameter reduction. Inspired by these innovations [17-18], our approach pivots towards employing depth-wise separable convolution layers.

Deep Neural Networks (DNNs), with a spotlight on CNNs, have garnered acclaim in Facial Emotion Detection [19,20]. Though deeper CNN models excel in varied image-processing tasks [27], current Visual Emotion Detection methodologies predominantly capitalize on the initial layers of CNNs. This likely stems from challenges inherent to emotion detection, including the prerequisite for high-resolution imagery and the intricacies of training profound CNNs with abundant hidden layers. The "vanishing gradient" conundrum further compounds these challenges [28].

The Residual Neural Network (ResNet) [29], pioneering the concept of skip-connection, has made waves in addressing the vanishing gradient dilemma. Its crux lies in fusing the input with the output post traversing several layers, enhancing layer information. Variants of this model such as ResNet-18, ResNet-34, ResNet-50, and ResNet-152 have made significant strides in the field.

DenseNet [30], another groundbreaking architecture, has introduced an intricate web of skip connections. Each layer in

***Article Received:** 12 August 2023 **Revised:** 22 September 2023 **Accepted:** 10 October 2023*

this paradigm is nourished by its antecedent layers and subsequently influences all subsequent ones. Its unique design, which amalgamates the input of a single layer with the concatenated output of preceding layers, mitigates the information bottleneck, optimizing computational efficiency. Our focal interest gravitates towards the DenseNet-121 variant, which is distinguished by its 120 convolutional layers nestled within four blocks.

Prominent deep CNN models, including VGG-16, ResNet-50, ResNet-152 [31], Inception-v3 [15], DenseNet121, and DenseNet-161 [32], stand testimony to the dynamic evolutions in the field. However, the onus of training these deep architectures necessitates a robust dataset and formidable computational prowess.

So, we have picked some pitfalls from previous related work and tried to cover few challenges among them which are as follows: (i) Minimize error rate and increase accuracy for frontal face as well as side face images (ii) Use of Pre-trained architectures to minimize the pre-training computation time and cost. (iii) To provide a Real-time efficient and deployable Emotion Detection System. The proposed work discussed further in next section.

III. PROPOSED WORK STRATEGY:

The proposed stream-lined workflow represented in the block diagram as in fig. 4 and 5. The steps in the workflow are outlined sequentially below. (i) Input frame extracted from real time device or external input source (datasets) (ii) Face Detection with pre-trained architectures (MobileNetV1-SSD, Tiny Yolo2) and Facial-Landmark model (with 68 view points) (ii) Visual Face Descriptors Extraction with VFDNet (Resnet- 31 layers) (iii) Visual Emotion Detection and Analysis with proposed VEDANet (baseline: Densenet-121) and extended the architecture by adding the OTO Layer in the end. The stepwise work progress is described below.

A. Face Detection:

In the realm of face detection, our primary focus is to achieve optimal accuracy in identifying face bounding boxes rather than merely ensuring quick inference times. To attain commendable accuracy for both static inputs and live frames from real-time devices, we amalgamate the strengths of mobilenet SSD and TinyYolo.

Our implementation employs the mobilenet SSD (Single Shot Multi-Box Detector) as a foundation. This pre-trained architecture, fortified by the TensorFlow object detection API, has been honed on the WIDERFACE dataset. Such a configuration enables the deep network to ascertain the coordinates of faces within a frame, subsequently delivering both the bounding box coordinates and their associated probability scores. In parallel, the Tiny Face Detector emerges as an efficient, real-time face detection solution that excels in web and mobile environments. Although its speed, compactness, and low resource consumption eclipse the SSD Mobilenet V1 face detector, it faces challenges with smaller faces. This detector has been trained on a specialized dataset comprising roughly 14K images, each meticulously annotated with bounding boxes. Further, its ability to predict bounding boxes encompassing facial feature points renders it synergistic with subsequent face landmark detection, often yielding superior results compared to solely using SSD Mobilenet V1. By recognizing the individual prowess of the Tiny Face Detector and SSD MobileNet, we devised a hybrid conjunction of the two. This blended approach is then integrated with a landmark recognition model to further accentuate facial features. Essentially, our model mirrors a compact iteration of TinyYoloV2 but diverges by adopting depth-wise separable convolutions in lieu of Yolo's traditional ones.

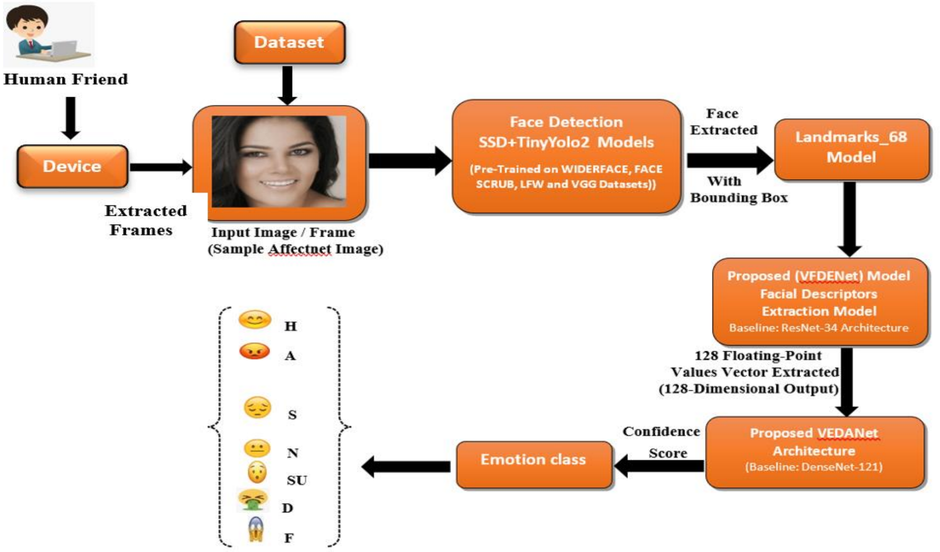

Human Friend
Dataset
Device
Extracted Frames
Input Image / Frame (Sample Affectnet Image)
Face Detection SSD+TinyYolo2 Models
(Pre-Trained on WIDERFACE, FACE SCRUB, LFW and VGG Datasets))
Face Extracted
With Bounding Box
Landmarks_68 Model
Proposed (VFDENet) Model Facial Descriptors Extraction Model
Baseline: ResNet-34 Architecture
128 Floating-Point Values Vector Extracted (128-Dimensional Output)
Proposed VEDANet Architecture
(Baseline: DenseNet-121)
Confidence Score
Emotion class
H
A
S
N
SU
D
F

Figure 4. Schematic Representation of the Proposed Visual Emotion Detection and Analysis Framework. The emotion classes, represented by H, A, S, N, SU, D, F, correspond to Happy, Angry, Sad, Neutral, Surprised, Disgust, and Fear, respectively.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10
***Article Received:** 12 August 2023 **Revised:** 22 September 2023 **Accepted:** 10 October 2023*

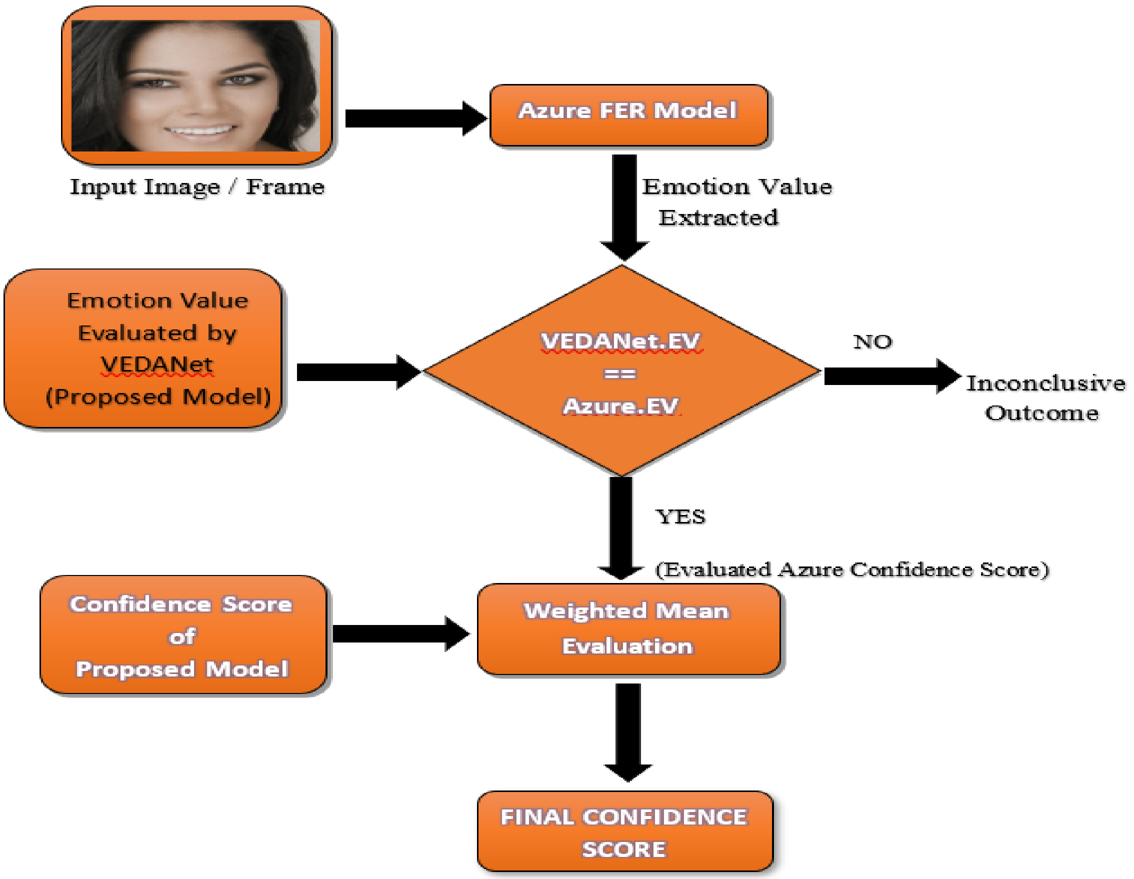


Figure 5. (Over-The-Top Optimization) OTO Layer Workflow to ensure the Highest Confidence Score and increase the Success Rate. *(Ensures FCS by comparing with Threshold value, t=0.6, if (FCS >= t) then evaluate conclusive Emotion

B. *Visual Facial Features Extraction:*

In this phase of our research, we focus on the extraction of precise facial landmarks consisting of 68 key points, ensuring a lightweight, rapid, and accurate process. We offer two models: the standard pretrained "face landmark-68 model," which is a mere 351 kilobytes in size, and the more compact "face landmark-68 tiny model," measuring approximately 82 kilobytes. Both models employ densely interconnected blocks and depth-wise separable convolutions. They were meticulously trained on a dataset comprising around 35,000 facial images meticulously annotated with 68 facial-landmark points.

Recent developments in deep convolutional neural networks have propelled us forward in the realm of unconstrained face identification. These advancements have enabled us to handle facial images under challenging conditions, including extreme angles, variable lighting conditions, and diverse resolutions. In our quest to further refine our capabilities in addressing these challenges, we have introduced the VFDENet. This architecture, boasting 31 convolutional layers, is built upon the ResNet as its foundational network. It represents a modified iteration of the ResNet-34 network, with three layers omitted and the number of filters per layer halved to optimize computational efficiency. The primary objective of the VFDENet architecture is to compute a comprehensive face descriptor—a feature vector comprising 128 values—for any given facial image. This descriptor serves as a means to encapsulate and convey the unique characteristics of an individual's face. For an illustrative visualization of the VFDENet architecture, please refer to Figure 2.

In essence, our work in this phase encompasses the development of lightweight yet highly effective facial landmark models, as well as the introduction of the VFDENet architecture to extract rich face descriptors, ultimately contributing to more robust facial feature analysis.

***Article Received:** 12 August 2023 **Revised:** 22 September 2023 **Accepted:** 10 October 2023*

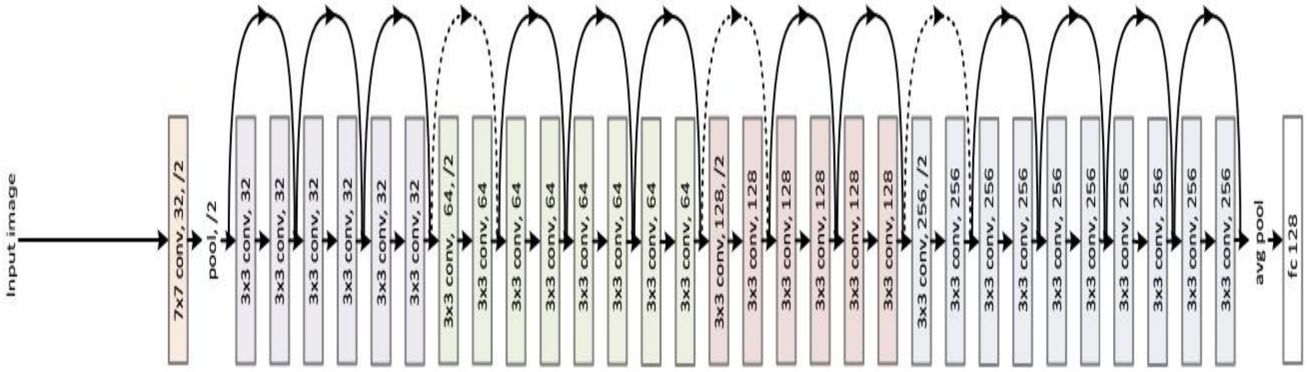


Figure 6. Face descriptor extraction using the ResNet-31 layers' VFDENet architecture. The output is a vector of 128 floating-point values, and the dotted shortcuts add more dimensions.

C. Visual Emotion Detection and Analysis with proposed VEDANet Architecture:

After Face Detection and Facial Features Extraction (Landmark points and descriptors), this stage will evaluate the Emotion Classification based on proposed model VEDANet architecture. Our proposed architecture exposed the exclusive outcomes comparative to SOTA due to the densely connected blocked network. The flow of working model is as follows in fig. 4-5 and the algorithms are shown in fig. 7. The algorithm representing multi-stage process for emotion classification and confidence assessment based on input data. Steps involved in fig. 7 algorithm are as follows :

1. **Initialization (Iteration X in Input):**
   - Input data in two forms: static images (Xss) and real-time frames (Xyo).
2. **Face Detection and Landmark Extraction:**
   - Apply SSD (Single Shot MultiBox Detector) to static images, obtaining face bounds (Fbo).
   - Apply YOLO (You Only Look Once) to real-time frames, obtaining face bounds (Fbo) and 68 facial landmarks (Lx).
3. **Feature Extraction:**
   - Extract face descriptors (Dface) from the facial landmarks (Lx).
   - Apply a Dense Neural Network to combine Dface with the original input (X) to obtain emotion class predictions (Ex).
4. **Emotion Classification:**
   - Convert the emotion class predictions (Ex) into an array (E).
5. **OTO Algorithm (Optimization Over the Top):**
   - For each iteration in the input:
   - Separate the input into two batches: User Sentiment (Xus) and Machine Sentiment (Xms).
   - Process both batches to obtain emotion and confidence pairs (XX).
   - Merge the results from the User Sentiment and Machine Sentiment batches (Eus, Ems).
6. **Confidence Calculation:**
   - If the User Sentiment (E1) matches the Machine Sentiment (E2):
   - Calculate the confidence as the weighted average of ConfE1 and ConfE2.
   - Set a threshold (T) for confidence.
   - If the calculated confidence is greater than the threshold, consider it a success.

TABLE I. EMOTION DETECTION (ED) ALGORITHM AND EMOTION ANALYSIS (EA) ALGORITHM

For iteration X in (input) do **X(static) >> SSD batch >> Xss** **X << Xss** **X(real frame) >> Yolo batch >> Xyo** **X<< Xyo** **X(Xss, Xyo) Input batch classifier** **Fbo << X (Xss, Xyo) Face bounds** **Lx << fo (Lmo). 68 landmarks** **Dface << fo (Rx)** **Ex << Dense(Dface(Rx), X). Emotion class** **E << arr(Ex)** **End**	***OTO algorithm-*** **For iteration X in (input) process** **X(us) >> POP batch >> XX** ***(XX-> emotion and confidence pair array)*** **X(ms) >> POP batch >> XX** **XX << POP batch** **XX >> (Eus, Ems)** **NOW IF** **E1 << Eus** **E2<< Ems** **E1==E2** **Do** **Conf<<ConfE1.ConfE2/attention-weighted average** **T=0.6** **Conf>T** **Success**

Consider an input image x_0 (128*128) that has been processed via the convolution layer. Proposed Architecture anatomy shown below in Table 1, that takes the input image and turned half the size and doubled the filters and reduced number of parameters. The first convo block has 64 filters of 8*8 size and 2 strides. *Convo-layer is made up of feature* *maps and filters. Where feature maps are the outcome as per the weight applied to the filter and filters hold the input weights in accordance with that output value.* As in our architecture, 1st convo-layer is 8*8 and every dense block consists of two-convolutions with 1*1 and 3*3 sized kernels. After the Convolution layers, there is a subsequent 3x3 Max Pooling layer with a stride of 2. The 'l-th' layer is characterized by having 'k0 + k × (l − 1)' input feature-maps, where 'k0' denotes the number of channels in the input layer. The convo layer and pooling layer together yield k feature-maps. k equals 32 in the current network design. The network growth rate is referred to as the hyper-parameter k. *To reduce sample feature detection in feature maps, pooling is necessary.* Pooling layers, by aggregating feature presence within map patches, provide an effective means of down sampling feature maps. In our architecture, we employ a batch normalization layer, followed by a 1x1 convolutional layer, and conclude with a 2x2 average pooling layer at each transition layer to facilitate down-sampling. In our implementation, we assumed that each 1x1 convolutional layer generates approximately 4k feature maps.

the pooling operation draws attention to the characteristics that overlap with filters by swiping a 2-D filter across the 3-D feature-map. The result of the pooling will be $((h - f + 1)/ s * (w - f + 1) * c)$, if a feature map with dimensions of w* h* c is provided. In this context, 'h' represents the height of the feature map, 'w' denotes its width, and 'c' signifies the channels within the feature map. Additionally, the filter's dimensions are expressed as 'f,' and the stride length is represented as 's.'

Additionally, the convolutional layer's most noticeable feature is used by the layer in max-pooling to operate on the feature map. The sharpest features on the image are the best lower-level features of the image, which can be extracted with the use of max-pooling. In proposed architecture, before the first dense block operations, a 3*3 max pooling operation is deployed to extract the most prominent feature from the feature map.

When using average pooling, the layer chooses the mean values of the items present in the feature map patch. The entire feature map is basically down-sampled to the average value recorded by the feature map's region. Each Transition Layer in the design uses 2*2 Average Pooling to assess the average value recorded by the region of the feature map.

So, the most apparent feature of any patch is thus revealed by max-pooling, whereas average pooling reveals the average of the covered region. Max and Average Pooling Operations represented as below in Fig. 7. After receiving the input, our convolutional blocks adhere to a structured sequence: Batch Normalization is first applied, followed by the ReLU activation function, and subsequently, a 2D Convolutional layer is employed. Our architecture is structured around four dense blocks, each consisting of two convolutional layers with 1x1 and 3x3 kernels. These convolutional operations are repeated within each block, with Block-1 undergoing 6 repetitions, Block-2 with 12 repetitions, Block-3 with 24 repetitions, and Block-4 with 16 repetitions.

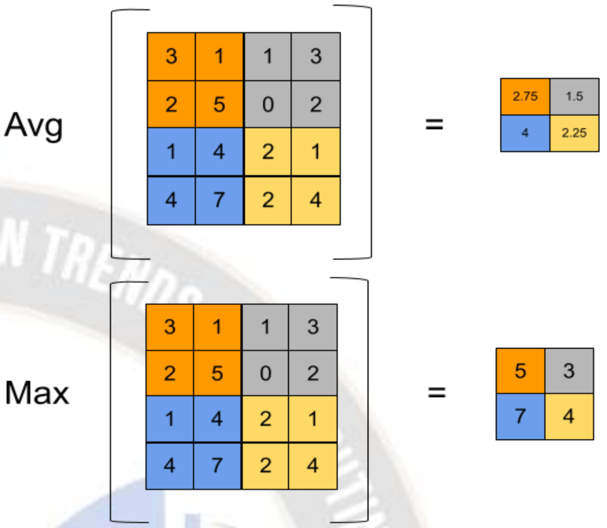


Figure 7. Illustration of Max Pooling and Average Pooling Operations

To mitigate potential overfitting caused by an increasing number of channels in the feature map, we introduce Transition Layers. These Transition Layers serve to halve the current channel count. They incorporate a 1x1 convolutional layer and a 2x2 Average Pooling Layer with 2-stride. We strategically stack Transition Layers between each dense block to maintain model compactness and efficiency.

Later, till block-4, the output is structurally compact enough and hence, it is directly passed to classification layer as it's represented in Table 1 (VEDANet Architecture). In proposed architecture, at *Classification Layer*, the global average pooling layer averages each feature map and global max-pooling-layer extract maximum from each feature map then send the resulting vector (concatenation of both) direct into the softmax layer, *avoiding the possibility of overfitting*. Final output Layer as: FC Layer provides the emotion classification with respect to the input image X_0 (128*128).

To further enhance our workflow for real-time live streams, we applied a hybrid approach to tap in potential of parallel processing, wherein we will parallelly assess our findings by taking advantage of another state-of-the-art cognitive service solution like, Microsoft Cognitive Services. In OTO layer we process the same input with Microsoft Cognitive Emotion Services and perform an attention-weighted average of the confidence scores to evaluate final confidence score (FCS). Further, we observed that if we tune the threshold hyperparameter to be equal to 0.6 (i.e. T=0.6) the workflow achieves higher accuracy while classifying sub level classification predictions as inconclusive in turn decreasing

Article Received: *12 August 2023* ***Revised:*** *22 September 2023* ***Accepted:*** *10 October 2023*

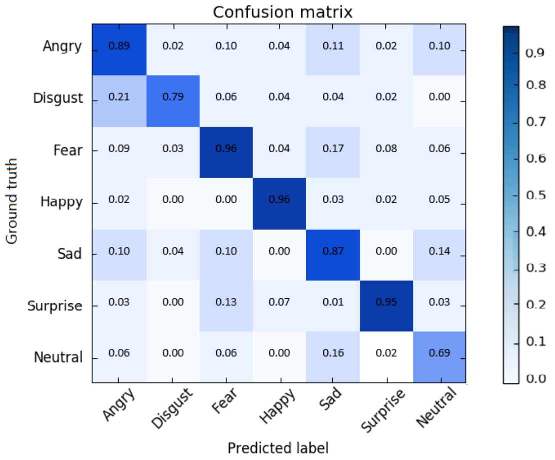
the error rate at greater extent (~4-5%). Finally, if the output of attention-weighted average is higher than the threshold (FCS >= T), we declare success. To assess the viability of our proposed model we also compared our model with SOTA benchmark datasets (AffectNet, FER2013, Yale, Google FEC) and Real- time live streaming.

TABLE II. VEDANET ARCHITECTURE (K = 32 GROWTH RATE). THE BN-RELU-CONV SEQUENCE IS APPLIED CONSISTENTLY TO EACH "CONV" LAYER AS OUTLINED IN THE TABLE.

Layers	**Output Size**	**VEDANet Architecture (K=32)** **(Description)**
Convolution	128x128	8x8 Convolution Layer (CL), Strides-2
Pooling	64x64	3x3 Max-Pooling, Strides-2
Dense Block (DB1)	64x64	1x1 Convolution
		3x3 Convolution (repeated 6 times)
Transition-Layer (1)	64x64	1x1 Convolution (CL)
	32x32	2x2 Average Pooling, Strides-2
Dense Block (DB2)	32x32	1x1 Convolution
		3x3 Convolution (repeated 12 times)
Transition-Layer (2)	32x32	1x1 Convolution (CL)
	16x16	2x2 Average Pooling, Strides-2
Dense Block (DB3)	16x16	1x1 Convolution
		3x3 Convolution (repeated 24 times)
Transition-Layer (3)	16x16	1x1 Convolution (CL)
	8x8	2x2 Average Pooling, Strides-2
Dense Block (DB4)	8x8	1x1 Convolution
		3x3 Convolution (repeated 16 times)
Classification Layer	1x1	Concatenation of 8x8 global average pooling & global max pooling
		FC Layer, Softmax

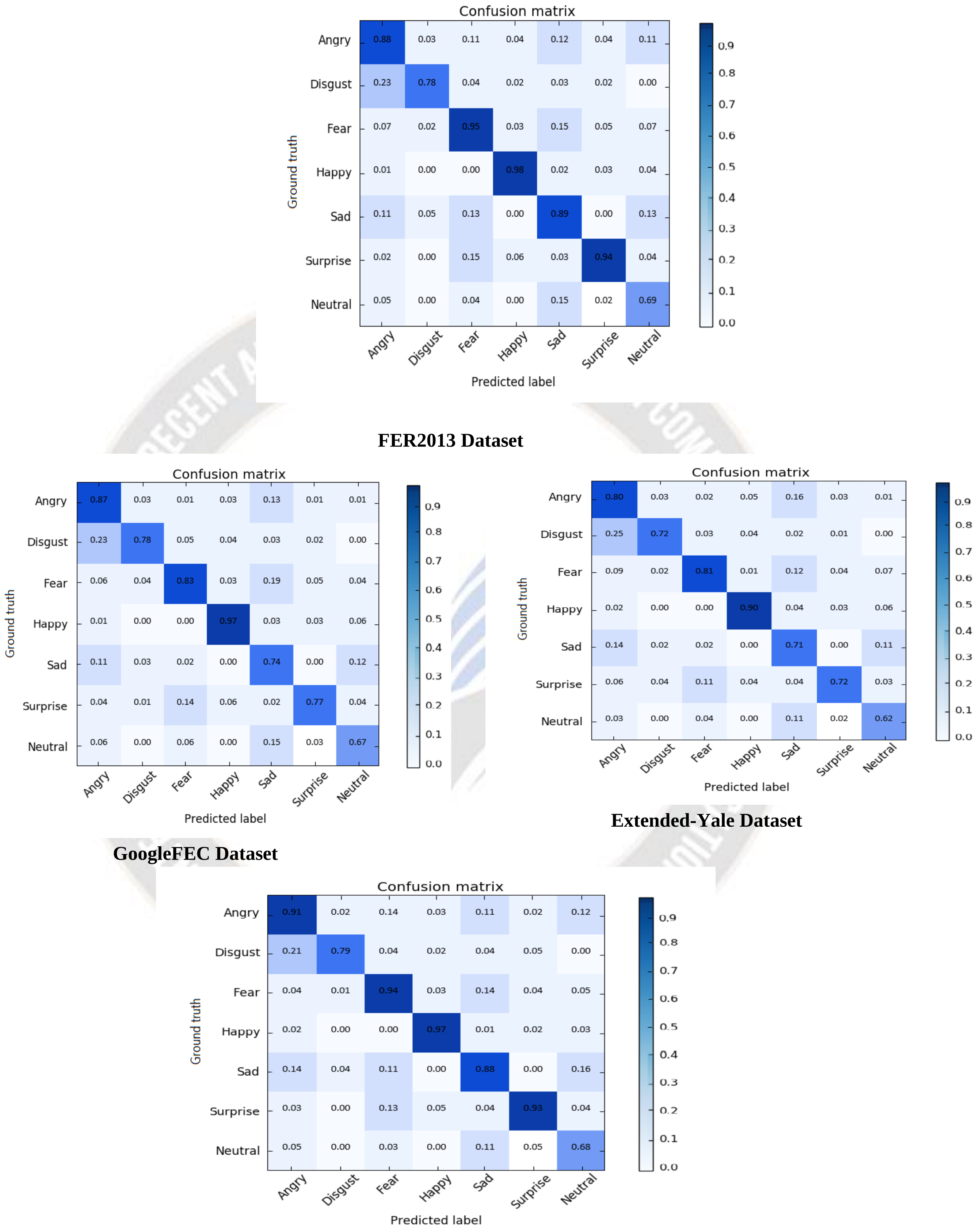
IV. PERFORMANCE EVALUATION AND COMPARATIVE STUDY

Proposed Model proved the accuracy 87.30% on AffectNet dataset, 92.76% on Google FEC, 95.23% on Extended Yale Dataset and 97.63% on FER2013 dataset and 98.34 % on Real-Time Dataset respectively. The Confusion Matrix with respect to each benchmark datasets shown below in figure 8. Table III also shows a comparative study of the proposed working-model with current best practices.



AffectNet Dataset

Article Received: *12 August 2023* ***Revised:*** *22 September 2023* ***Accepted:*** *10 October 2023*



FER2013 Dataset

Extended-Yale Dataset

GoogleFEC Dataset

Real-Time Dataset

Figure 8. Confusion Metrics for Emotion Classes (7 EC) with respect to considered dataset

Article Received: *12 August 2023* ***Revised:*** *22 September 2023* ***Accepted:*** *10 October 2023*

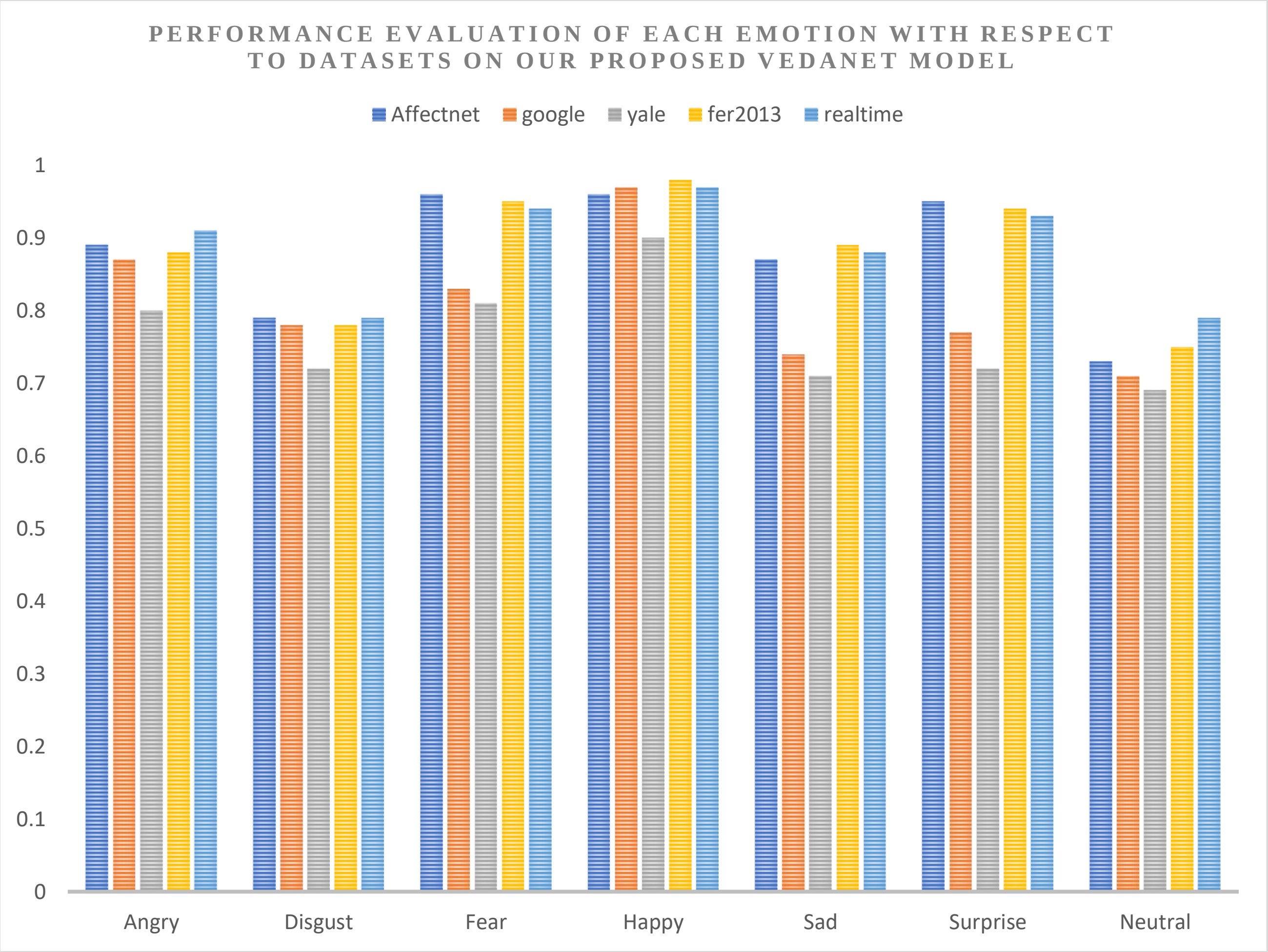


Figure 9. Comparative Analysis of Evaluated Outcomes (Emotion Classes (7 EC) with respect to considered datasets)

TABLE III. COMPARATIVE STUDY OF PERFORMANCE EVALUATION OF PROPOSED VEDANET ARCHITECTURE WITH STATE-OF-THE ARTS

ARCHITECTURE	ACCURACY	DATASET USED
CNN [33]	89.98%	FER2013
ER_MOPI (META-LEARNING) [34]	ACC: 90% (MULTI-PIE) ACC: 68% (AFFECTNET DATASET)	AFFECTNET, CMU MULTI-PIE
DENSESANET121 [35]	60.88%	AFFECTNET
HOG+VGG-FACE FEATURE-FUSION MODEL [36]	95.26% (5-FOLD-CROSS VALIDATION)	YALE-FACE
DISTILLED CNN [37]	61.6%, 86.5%	AFFECTNET, GOOGLE FEC
TRANSFER LEARNING WITH CNN [38]	RMSE: 0.09 - (VALENCE) RMSE: 0.1-(AROUSAL)	AFFECTNET AND AMIGOS
VEDANET (OURS)	97.63%	FER-2013
VEDANET (OURS)	95.23%	YALE-FACE
VEDANET (OURS)	92.76%	GOOGLE FEC
VEDANET (OURS)	87.30%	AFFECTNET
VEDANET (OURS)	98.34%	REAL-TIME (LIVE FRAMES)

V. CONCLUSION AND POSSIBLE FUTURE DIRECTIONS:

For the purpose of visual emotion detection and analysis from facial frames (static or live frames), an effective VEDANet architecture has been proposed in this study.

In the proposed model, the application of *Transfer Learning technique used* for feature extraction from facial frames. Based on the experimental findings, hybrid pre-trained Deep Convolutional Neural Network (DCNN) models were applied to renowned emotion datasets, including FER2013, AffectNet, Google FEC, and Yale. The proposed framework has extremely high recognition accuracy. For the sake of simplicity and a few perplexing facial photos, trials carried out in normal circumstances without using the pre-trained DCNN model were misclassified in the current study. The classification accuracy saw enhancements through fine-tuning the hyperparameters of each pre-trained model, incorporating the Top Optimization Layer (OTO-Layer), and closely analyzing facial expressions within the frames. When compared to cutting-edge approaches, the suggested model performs effectively and achieved the accuracy of 87.34% on AffectNet dataset, 92.76% on Google FEC, 95.23% on Yale Dataset and 97.63% on FER2013 dataset and 98.34% on Real-Time live frames. Broader real-world commercial applications, including patient monitoring in hospitals or security surveillance, will be suitable with the presented research work. Additionally, to address new industrial applications, the concept of face emotion detection and analysis may be expanded to include emotion recognition from body language or vocal media (speech).

REFERENCES:

1. Sariyanidi, E., Gunes, H., & Cavallaro, A. (2014). Automatic analysis of facial affect: A survey of registration, representation, and recognition. *IEEE transactions on pattern analysis and machine intelligence*, *37*(6), 1113-1133.
2. Anagnostopoulos, C. N., Iliou, T., & Giannoukos, I. (2015). Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011. *Artificial Intelligence Review*, *43*(2), 155-177.
3. Marechal, C., Mikolajewski, D., Tyburek, K., Prokopowicz, P., Bougueroua, L., Ancourt, C., & Wegrzyn-Wolska, K. (2019). Survey on AI-Based Multimodal Methods for Emotion Detection. *High-performance modelling and simulation for big data applications*, *11400*, 307-324.
4. Alkawaz, M. H., Mohamad, D., Basori, A. H., & Saba, T. (2015). Blend shape interpolation and FACS for realistic avatar. *3D Research*, *6*(1), 1-10.
5. Shan, G. (2009). McOwan, 2009 Shan C., Gong S., McOwan PW. *Facial expression recognition based on Local Binary Patterns: A comprehensive study, Image and Vision Computing*, *27*(6), 803-816.
6. Zhang, S., Li, L., & Zhao, Z. (2012, October). Facial expression recognition based on Gabor wavelets and sparse representation. In *2012 IEEE 11th international conference on signal processing* (Vol. 2, pp. 816-819). IEEE.
7. Rouast, P. V., Adam, M. T., & Chiong, R. (2019). Deep learning for human affect recognition: Insights and new developments. *IEEE Transactions on Affective Computing*, *12*(2), 524-543.
8. Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR).
9. Liu, Z. T., Li, S. H., Cao, W. H., Li, D. Y., Hao, M., & Zhang, R. (2019). Combining 2D gabor and local binary pattern for facial expression recognition using extreme learning machine. *Journal of Advanced Computational Intelligence and Intelligent Informatics*, *23*(3), 444-455.
10. Deeb, H., Sarangi, A., Mishra, D., & Sarangi, S. K. (2022). Human facial emotion recognition using improved black hole based extreme learning machine. *Multimedia Tools and Applications*, 1-24.
11. B+. Available online: https://computervisiononline.com/dataset/1105138686 (accessed on 29 November 2017).
12. Vemulapalli, R., & Agarwala, A. (2019). A compact embedding for facial expression similarity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 5683-5692).
13. Mollahosseini, A., Hasani, B., & Mahoor, M. H. (2017). Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, *10*(1), 18-31.
14. Goodfellow, I. J., Erhan, D., Carrier, P. L., Courville, A., Mirza, M., Hamner, B., ... & Bengio, Y. (2013, November). Challenges in representation learning: A report on three machine learning contests. In *International conference on neural information processing* (pp. 117-124). Springer, Berlin, Heidelberg.
15. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2818-2826).
16. He, K., Zhang, X., Ren, S., & Sun, J. (2015). Deep residual learningfor image recognition. *ComputerScience*.
17. Huang, G., & Liu, Z. (2017). vd Maaten L., Weinberger KQ,“. In *Densely connected convolutional networks,” in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 2261-2269).
18. Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1251-1258).
19. Alom, M.Z., Taha, T.M., Yakopcic, C., Westberg, S., Sidike, P., Nasrin, M.S., Hasan, M., Van Essen, B.C., Awwal, A.A. and Asari, V.K., (2019). A state-of-the-art survey on deep learning theory and architectures. *Electronics*, *8*(3), p.292.
20. Sahu, M., & Dash, R. (2021). A survey on deep learning: convolution neural network (CNN). In *Intelligent and Cloud Computing* (pp. 317-325). Springer, Singapore.
21. Mollahosseini, A., Chan, D., & Mahoor, M. H. (2016, March). Going deeper in facial expression recognition using deep neural networks. In *2016 IEEE Winter conference on applications of computer vision (WACV)* (pp. 1-10). IEEE.
22. Zhao, X., Shi, X., & Zhang, S. (2015). Facial expression recognition via deep learning. *IETE technical review*, *32*(5), 347-355.
23. Li, J., Huang, S., Zhang, X., Fu, X., Chang, C. C., Tang, Z., & Luo, Z. (2018, December). Facial expression recognition by transfer learning for small datasets. In *International Conference on Security with Intelligent Computing and Big-data Services* (pp. 756-770). Springer, Cham.
24. Bendjillali, R. I., Beladgham, M., Merit, K., & Taleb-Ahmed, A. (2019). Improved facial expression recognition based on DWT feature for deep CNN. *Electronics*, *8*(3), 324.
25. Ngoc, Q. T., Lee, S., & Song, B. C. (2020). Facial landmark-based emotion recognition via directed graph neural network. *Electronics*, *9*(5), 764.
26. Pranav, E., Kamal, S., Chandran, C. S., & Supriya, M. H. (2020, March). Facial emotion recognition using deep convolutional

neural network. In *2020 6th International conference on advanced computing and communication Systems (ICACCS)* (pp. 317-320). IEEE.

27. Khan, A., Sohail, A., Zahoora, U., & Qureshi, A. S. (2020). A survey of the recent architectures of deep convolutional neural networks. *Artificial intelligence review*, *53*(8), 5455-5516.
28. Hochreiter, S., Bengio, Y., Frasconi, P., & Schmidhuber, J. (2001). Gradient flow in recurrent nets: the difficulty of learning long-term dependencies.
29. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770-778).
30. Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4700-4708).
31. Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.
32. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., ... & Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1-9).
33. Zhou, F., Kong, S., Fowlkes, C. C., Chen, T., & Lei, B. (2020). Fine-grained facial expression analysis using dimensional emotion model. *Neurocomputing*, *392*, 38-49.
34. Kuruvayil, S., & Palaniswamy, S. (2021). Emotion recognition from facial images with simultaneous occlusion, pose and illumination variations using meta-learning. *Journal of King Saud University-Computer and Information Sciences*.
35. Wang, C., Lu, K., Xue, J., & Yan, Y. (2019). Dense attention network for facial expression recognition in the wild. In *Proceedings of the ACM Multimedia Asia* (pp. 1-6).
36. Ahadit, A. B., & Jatoth, R. K. (2022). A novel multi-feature fusion deep neural network using HOG and VGG-Face for facial expression classification. *Machine Vision and Applications*, *33*(4), 1-23.
37. Schoneveld, L., Othmani, A., & Abdelkawy, H. (2021). Leveraging recent advances in deep learning for audio-visual emotion recognition. *Pattern Recognition Letters*, *146*, 1-7.
38. Rescigno, M., Spezialetti, M., & Rossi, S. (2020). Personalized models for facial emotion recognition through transfer learning. *Multimedia Tools and Applications*, *79*(47), 35811-35828.