IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

Decoding Human Facial Emotions: A Ranking Approach using Explainable AI

SUDHEER BABU PUNURI[1], SANJAY KUMAR KUANAR[2], TUSAR KANTI MISHRA[3](MEMBER, IEEE), VEERANKI VENKATA RAMA MAHESWARA RAO[4], AND SHIVA SHANKAR REDDY[5],

[1, 2]Department of Computer Science and Engineering, GIET University, Gunupur, Odisha, India.
[3]Department of Computer Science and Engineering, Manipal Institute of Technology Bengaluru, Manipal Academy of Higher Education, Manipal, Karnataka, India - 576104 (e-mail: tusar.mishra@manipal.edu)
[4]Department of Computer Science and Engineering, Shri Vishnu Engineering College for Women, Bhimavaram, India.
[5]Department of Computer Science and Engineering, Sagi Ramakrishnam Raju Engineering College (A), Bhimavaram, Andhra Pradesh, India.

Corresponding author: Tusar Kanti Mishra (e-mail: tusar.mishra@ manipal.edu).

ABSTRACT To decipher human activities and facilitate organic computer-human interactions, facial expression recognition is crucial using four datasets JAFFE, CKPlus, KDEF and AffectNet, we achieve excellent accuracy in human face emotion classification with the VGG 16 pre-trained model with transfer learning. To obtain a comprehensive insight of the model's decision-making process, we employ Layerwise Relevance Propagation (LRP), a method from explainable Artificial Intelligence (XAI). Only positive relevance scores are taken into account for successfully predicted test images from the datasets. Contributory pixels towards predicting the intensity of emotion are pixels with good relevance ratings. By combining emotion recognition with LRP, we can forecast emotion labels and ranks. Using a confusion matrix, we checked if our predictions were in line with reality. Our model achieved intensity prediction accuracy of 96.33% on JAFFE, 95.78% on CKPlus, 95.78% on KDEF, and 93.89% on AffectNet. A group of ten annotators work together to generate ground truth by assigning ratings of "MINIMAL, "AVERAGE, and "STRONG to each image. This study demonstrates how well our method predicts the ranks of emotion intensity and provides information on how trustworthy and interpretable the model is. Facial emotion recognition's intensity ranking is made more robust with the addition of XAI techniques like LRP.

INDEX TERMS Explainable Artificial Intelligence, Facial Emotion Recognition, Layer-wise Relevance Propagation, Ranks for Facial Emotions.

I. INTRODUCTION

The goal of the scientific and technological discipline known as facial emotion recognition (FER) [1] is to identify, understand, and communicate human emotions using facial expressions. To automatically collect and evaluate facial data for emotion classification, FER systems employ sophisticated algorithms. Smart homes [2], healthcare [3], education [4], automotive, virtual reality/augmented reality (VR/AR), and social schemes [5] are just a few of the many uses for FER. The six most common facial expressions that FER models identify are surprise, wrath, contempt, fear, and happiness or sadness [6]. Some datasets also contain neutral expressions and contemptuous ones. Facial expression intensity (PEI) refers to the level or scale of an emotion conveyed by human facial expressions. A person's facial expressions, like a look of happiness or fear, can be measured by this. Emotion detection relies on accurate PEI capture and interpretation since it provides more insight into human emotions and reactions. In 2014, VGG16 [7], a CNN architecture, was introduced by the Visual Geometry Group at the University of Oxford. Thirteen convolutional layers and three fully linked layers make up VGG16's sixteen layers. The layout is uniform and uncomplicated. To reduce the number of spatial dimensions, it employs max-pooling layers with 2x2 windows, 3x3 filters activated by ReLU, and 3x3 filters. VGG16 is a popular choice for transfer learning since it is usually used for picture classification and is pre-trained on ImageNet.This algorithm has been extensively applied in computer vision tasks such as object identification and segmentation of images because of its adaptability, easiness, and strong feature learning capabilities. A suitable technique in the fields of deep learning and machine learning, transfer learning [8] involves taking the information learned in one task and transferring it to another closely related one. Transfer learning reduces the

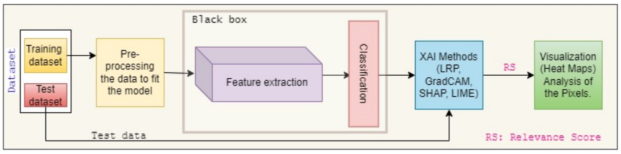

Black box
Dataset
Training dataset
Test dataset
Pre-processing the data to fit the model
Feature extraction
Classification
XAI Methods (LRP, GradCAM, SHAP, LIME)
RS
Visualization (Heat Maps) Analysis of the Pixels.
Test data
RS: Relevance Score

FIGURE 1. A simple Neural Network structure.

requirement for large amounts of training data and computational resources by utilizing pre-trained models that have been trained on extensive datasets when approaching new tasks. These pre-built models have the ability to retrieve important information and context that is useful for the task at hand, thanks to their learned characteristics and representations. Using this method, the training process is accelerated, and the model's performance and generalizability are enhanced, even when the target dataset is small. One area of AI that is gaining popularity is Explainable Artificial Intelligence (XAI) [9] [10] [11] [12] [13] which aims to provide logical explanations for the decisions and actions used by AI systems. Our study emphasizes the significance of XAI in making AI models more transparent and readable, particularly for emotion recognition. The purpose of the field called explainable artificial intelligence (XAI) is to make black box [14] models more understandable. It is possible to detect biases in black box models using XAI techniques, and these models can make decisions. By incorporating XAI [15] into FER, we can better understand how the model makes decisions and identify biases that could lead to inaccurate predictions. To ensure reliable emotion recognition, our FER system leverages XAI's visualization power. By generating heatmaps, we pinpoint exactly which pixels in an image hold the most weight in determining the predicted emotion, leading to more accurate and trustworthy results.

When it comes to evaluating mental health and Human-Computer Interaction (HCI), it is crucial to be able to explain the logic behind AI-powered emotion detection to build trust and acceptance among users. Furthermore, examining the inner workings of Explainable Artificial Intelligence (XAI) ensures that expert validation, system integrity, and fairness are maintained in AI models. With the advent of AI-driven solutions, the use of XAI in facial expression recognition promotes ethical and responsible AI development, paving the way for transparent, reliable, and unbiased decision-making in real-world settings. Fig. 3 illustrates the overall method of XAI. Many researchers in the field of explainable artificial intelligence (XAI) use a method called layer-wise relevance propagation (LRP) [16] [17] [18] [19]. LRP rates the relevance of specific input parameters to focus on the pixels contributed during prediction process of a deep learning model. Researchers and practitioners can use these relevance scores to see how important each feature was in the model's final prediction, which helps them understand which input data items affected the model's decision-making. The core idea behind LRP is to recursively distribute the significance of the model's output among its layers according to the proportion of activation that each neuron contributes to the output.

Motivation: The exploration of human emotions is a potential field for a myriad of directions. Human-computer interaction: Reactions to emotions can enhance our relationship with virtual assistants or interactive technologies. This could help create a safer and more productive human-robot collaboration, that necessarily requires better robotic capabilities to understand the intensity of the emotions of the person visiting. Hence security and surveillance can profit from emotion analysis to recognize any likely danger or unexpected conduct. Key Finding: Adding emotion to the equation improves biometric recognition. Now, these new advancements in biometric recognition systems added a twist: emotion. Emotion intensity has been adopted as one of the indicative indices in healthcare and measures the intensity of emotions and provides vital information on mental health conditions that are used for the customization of patient care. This emerging research field potentially places conclusive demand on a deep knowledge of human psychology, traditional flavors, and emotions, as well as becoming promising for technological application as well as for human well-being. To summarise, the main contribution of the research is:

1) We employ Transfer Learning (TL) to construct a proficient Deep Convolutional Neural Network (DCNN) model.
2) Utilizing layer-wise relevance propagation (LRP) to calculate the relevance scores of the involved pixels.
3) Based on the obtained relevance scores, assign "MINIMAL", "AVERAGE", and "STRONG" ranks to the images.
4) Validate the effectiveness and robustness of our approach using performance metrics on lab-controlled datasets (CKPlus, JAFFE, KDEF) as well as on the AffectNet dataset, which represents more unconstrained, real-world scenarios.

The following sections are structured in the following way: Facial Emotion Recognition (FER), Explainable Artificial Intelligence (XAI), Layer-wise Relevance Propagation (LRP), and Intensity Estimation are some of the subjects covered in Section 2's literature analysis. Section 3, named "Methodology," describes the model in detail, including its architecture, training techniques, validation and assessment metrics, and layer-wise relevance propagation. This section also elaborates on the process of ranking emotions based on LRP relevance scores. Moving on to Section 4, we now focus on revealing findings and sparking conversation. This involves giving a brief summary of data sets, displaying a confusion matrix relevant to Facial Emotion Recognition (FER), creating heatmaps, evaluating emotions through LRP, performing initial experiments for validation, confirming rankings obtained from relevance scores, contrasting findings with previous studies, and participating in detailed conversations. In conclusion, Section 5 of the paper summarizes the findings

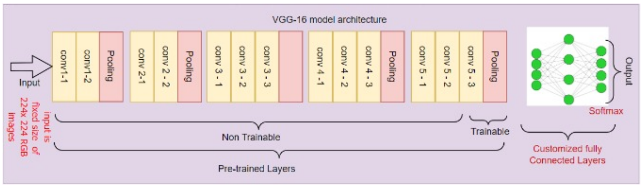


FIGURE 2. Magnetization as a function of applied field. VGG - 16 Architecture

and outlines possible directions for future research.

II. LITERATURE REVIEW

Facial emotion recognition holds significant potential across various applications, including the detection of drowsy drivers [20], early-stage abnormality identification in Alzheimer's disease and the development of crime prediction systems. The field of Computer Vision has witnessed a surge in popularity and effectiveness, largely attributable to the widespread adoption of transfer learning. Transfer learning has proven invaluable to researchers by not only saving time but also conserving computational resources. In a recent study authors [21] proposed a novel technique for FER using transfer learning based on PathNet was introduced. This approach aimed to address issues encountered in traditional transfer learning techniques. Demonstrating the efficacy of effective fine-tuning of pre-trained models, Ikromovich et al. [22] showed that it leads to improved accuracy and efficiency. The transfer learning applications in FER has extended into the healthcare domain. Notably, Theerthagiri et al. [23] experimented on the use of Deep Belief Networks and transfer learning to estimate emotion levels, specifically focusing on stress identification. The progress in computer vision has facilitated the presentation of numerous studies focused on extracting and classifying facial features through diverse facial expression techniques. This achievement is a testament to the continuous advancements in both research and technology. Significantly, a new framework for identifying and amplifying facial emotions through weighted voting has been presented, according to Kamarol et al. [24]. By making revisions to its upper dense layers, Helaly et al. [25], a different team of researchers enhanced ResNet. Sun et al [26] have undertaken a modification in manual facial emotion intensity, adopting the methodology presented by Shinohara et al. [27]. These endeavors collectively showcase the dynamic landscape of research in facial emotion recognition and intensity, reflecting the continuous evolution of methodologies and techniques in this domain. Following the initial stages of image sequencing, pre-processing, and feature extraction and selection, akin to the training phase, the extracted features underwent evaluation through the trained model. The determination of emotion intensity was then established based on the identified Action Units (AUs). The purpose of this comparative study is to lay the groundwork for our analysis by comparing various approaches employed at the feature extraction intermediate step and the classification critical step. Numerous studies on the intensity of facial emotions have focused on facial action units (FAUs).

These AUs were first proposed by Paul Ekman and Friesen, and without them, the repertoire of facial expressions is incomplete. Another term for them is facial landmarks. For instance, Action Unit 1 (AU1) corresponds to the inner brow raised, exemplifying the detailed and nuanced nature of facial action units in capturing and expressing emotion. Action Units (AUs) also contribute to generating the necessary information for evaluating and scrutinizing the intensity of emotions. The concept of class co-occurrences involves a set of AUs associated with a specific emotion class; for example, the combination of AU6, AU12, and AU25 constitutes the emotion "HAPPY". In a study by Shinohara et al. [27] 16 facial features derived from facial landmarks were suggested. However, there are limitations in the scope of intensity estimation due to the spontaneous nature of facial expressions. Some action units, such as AU15, where the corners of the lips point upwards and are angled towards the ground, present instances that can be misleading and non-additive. These challenges complicate the accurate determination of both emotion and its intensity based solely on AUs, impeding further progress in this context. The importance of understanding and trusting the decision-making processes of complex models has led to a significant focus on the interpretability of deep neural networks in recent years. Layerwise Relevance Propagation (LRP) is a method that has been suggested to explain the predictions made by deep neural networks. Gunning et al. [28] define LRP as a subset of eXplainable Artificial Intelligence (XAI). The idea of using LRP to interpret decisions generated by deep neural networks was first introduced by Bach et al. [16]. A research article by Samek et al. [29] delves into visualization methods aimed at interpreting the knowledge acquired by deep neural networks. This work evaluates visual interpretations, providing an understanding of the model's decision-making and advocating for transparency in AI systems. Addressing the challenges of interpretation for indistinguishable inputs and the robustness of interpretations, Ghorbani et al. [30] have proposed methodologies in this context. According to Bach et al., [16], using saliency maps can harness the power of explainability in deep learning models, removing the necessity for a subjective visual comparison of contour lines. These advancements collectively contribute to enhancing the interpretability and transparency of deep neural networks. Authors [31] introduced several lightweight models based on architectures such as MobileViT, MobileFaceNet, EfficientNet, and DDAMFN, which were trained in multi-task scenarios to recognize various aspects including facial expressions, valence, and arousal from static photos. Researchers [32] propose a multi-modal architecture for estimating Emotional Reaction Intensity (ERI) from videos. This approach combines video and audio information to measure the intensity of subjects' reactions to stimuli along various emotional dimensions.

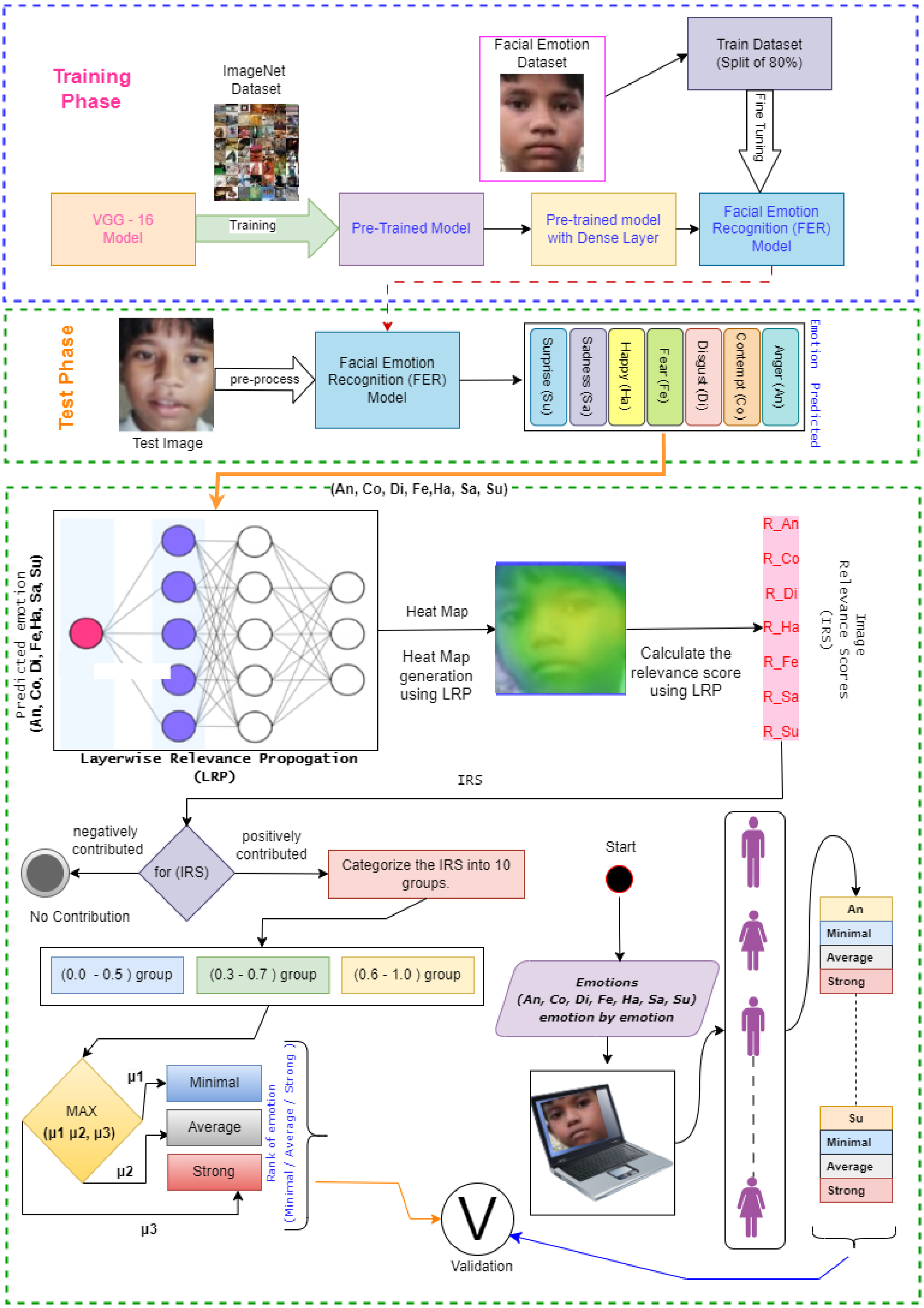


FIGURE 3. shows the proposed model architecture designed for classifying human emotions and assigning them Minimal, Average, or Strong labels based on their intensity

III. METHODOLOGY

Our method employs VGG-16 to classify facial expressions and evaluate emotions using four datasets: JAFFE, CKPlus, KDEF and AffectNet. The use of transfer learning in a deep learning architecture allows for the classification of facial expressions into seven predetermined emotions. Fine-tuning is applied to a pre-trained Convolutional Neural Network (CNN) to enhance the model's performance. Additionally, our research uses LRP to assess the significance of different facial features in detecting each emotion. A pilot study is underway to confirm the accuracy of our emotion ranking results by validating the annotations of emotion intensity. Our primary goal is to create a clear and consistent system for evaluating and examining the suggested plan. By utilizing LRP and fine-tuned CNNs, we aim to enhance the understanding of facial expression recognition in emotion classification through a thorough methodology.

A. MODEL ARCHITECTURE

VGG-16, as highlighted in a survey by Khan et al. [7], stands out as one of the highly favored convolutional neural network (CNN) architectures in recent times. Originating in 2014, it was meticulously crafted by Simonyan and Zisserman, and its design has since gained prominence in various applications, showcasing its enduring relevance and effectiveness in the field of deep learning. The VGG-16 model is a sophisticated deep convolutional neural network that comprises a grand total of sixteen layers. The initial thirteen layers include convolutional layers, then 3 fully connected layers follow. Every convolutional layer utilizes compact 3x3 filters, with additional filters incorporated in subsequent layers to enhance the model's capacity to identify intricate features. Max-pooling layers are strategically added within the convolutional layers to decrease spatial dimensions and abstract features even more. After the convolution layers, the model ends up with three dense layers that serve as the classification heads of the network. The final complexity provides the final result, and determines the possibility of facial expressions. In our study, we used transfer learning as the basis of Facial Emotion Recognition (FER) using the VGG-16 paradigm. We modified the model for our specific task by delaying the initial sequence and optimizing the final sequence, using extensive pre-trained knowledge from the large dataset. The choice to incorporate a Dense Layer with seven neurons into the fully linked layer was made with consideration for the seven emotion classes present in the FER datasets. By going this way, we were able to utilize the established capabilities of the VGG-16 architecture while also customizing it to the needs of our emotion recognition task. We created a new classifier on top of the pre-trained convolutional layers of the VGG-16 model to fit our emotion rating goal. This tailor-made classifier employed three fully connected layers in addition to adaptive average pooling. ReLU was used as the activation function in the first two fully connected layers, which had 4,096 neurons each. Dropout regularization with a rate of 0.5 was used to reduce overfitting after every ReLU activation. The ultimate fully connected layer included seven output neurons, matching with the desired number of emotion classes for prediction. By adding this extra classifier, we managed to obtain advanced visual features from the pre-trained VGG-16 model without unfreezing the convolutional layers. This architectural modification enhanced our classifier's ability to accurately recognize emotions, which was beneficial in our experiments evaluating emotion ratings. Our proposed framework mainly includes two stages. During the initial phase, the model is trained with VGG-16 by utilizing the JAFFE, CKPlus, KDEF and AffectNet datasets. After that, the model is evaluated using a test dataset, and the confusion matrix shows how accurate the predictions are. During the second phase, the model uses images that it accurately predicted to calculate relevance scores and create heatmaps through Layer-wise Relevance Propagation (LRP). The model assigns intensity rankings—Minimal, Average, and Strong—to the test sample images. To ensure the robustness of these assigned labels, a validation process is conducted. This involves comparing the model-assigned labels with those determined during a pilot experiment. The consistency between the model's labels and the pilot experiment's labels demonstrates the robustness and reliability of the model.

TABLE 1. Hyperparameters used during training the model

Training parameters	Value(s)
Optimizer	AdamW
Batch Size	32 or 64
Weight Decay	0.001
Loss Function	Categorical_Cross_ entropy
Number of training iterations (Epochs)	100 or 150
Batch Size	32 & 64
Dropout Rate	50
Regularization	0.001
Input image shape	$(224, 224, 3)$

B. CONFIGURATION FOR COMPUTATION

Experimentation took place on the Google Colab platform using a Tesla T4 GPU equipped with 9.063MiB memory. Deep learning tasks were carried out using Python version 3.10.6 and PyTorch version 2.0.1+cu118. Taking advantage of these computational resources helped to carry out training and evaluation tasks for our emotion ranking model efficiently, guaranteeing reproducibility and speeding up our research efforts.

C. DATA PRE-PROCESSING

Before training the model, data pre-processing is needed to be done. As a part of pre-processing two changes were made to the input images. At first, the images were adjusted to a consistent size of 255 x 255 pixels to guarantee uniformity in dimensions. Afterward, the images were changed to a tensor style, allowing smooth handling by the neural network. These initial processing stages standardize the input information, improving the model's ability to learn features effectively when training.

D. DATASET SPLIT FOR TRAINING, VALIDATION, AND TESTING

To conduct our experiments, we selected the CK+, JAFFE, and KDEF datasets. Our methodology entailed a deliberate partitioning of these datasets to facilitate effective training, validation, and evaluation of our emotion ranking model. In order to start the experimentation, we split the complete dataset into two parts: one for training and one for testing, maintaining an 80-20 percent ratio. 80 percent of the data in the training set was used for model training, and the remaining 20 percent was set aside for testing. Again, the training split is further splitted into training and validation split. The test split ensured that the model's accuracy was tested on new data in the stage. The validation split, helps to recognize and solve problems with overfitting for every epoch during training phase. Throughout the experimentation, the test set remained entirely separate, serving as an independent benchmark to evaluate the model's generalization capability and its ability to accurately rank emotions on instances not encountered during training. This dataset split methodology enabled thorough evaluations while ensuring the reliability and robustness of the model across various emotional instances.

E. TRAINING PROCEDURE

The VGG-16 model processed the pre-processed facial images using the training set, which also served to monitor the model's progress. Through iterative updates of the model's parameters using backpropagation, we aimed to minimize training loss and enhance performance, as detailed in Table 1. To gauge how effectively the model adapted to the training loss throughout the training process. To prevent overfitting and assess the model's generalization ability, we utilized the validation set, evaluating the model after each training session. This evaluation provided insights into validation correctness, offering a means to monitor the model's performance on previously unseen data. Such careful validation checks were crucial to ensure that the model accurately captured underlying patterns without relying excessively on the training data.

F. PROPOGATION OF RELEVANCE ACROSS LAYERS

In this part, we present the LRP method developed by Bache et al. [16]. The LRP algorithm acts as a tool for comprehending how neural networks function. It functions by tracking the involvement of every input node in the end output node, moving through the layers sequentially. A function of relevance is established, which receives the i_{th} pixel of image x as its input. When an image x is given, it produces a real number represented as $Rel\_S(x_i)$ which is called as the relevance score. In this context, x signifies the influence on the determination (outcome) $f(x)$, which may be positive or negative.

Constraint 1. If $Rel\_S(x_i) > 0$ it implies a beneficial (positive) contribution, where $Rel\_S(x_i) < 0$ indicates a waste (non-beneficial contribution or negative). If the range of Rel is limited to a non-negative number, $Rel\_S(x_i) = 0$ indicates no contribution, and $Rel\_S(x_i) > 0$ indicates a positive contribution.

Constraint 2.

$$f(x) = \sum_{i=1}^{V} Rel\_S(X_i) \tag{1}$$

In our framework, by adding Constraint (1) we quantify the rate of contribution of individual pixel x_i on $f(x)$. Within our model, when including Constraint(1), we assess the specific impact rate of each pixel x_i on $f(x)$. In our model, when we introduce Constraint (1), we measure how much each pixel contributes to LRP. As its name suggests, LRP utilizes a forward-feeding network architecture of a Deep Neural Network (DNN), specifically VGG-16 in our scenario. The first layer accepts pixel inputs, while the final layer outputs the classifier f''s prediction. LRP suggests that the relevance scores $Rel\_S(X_i)$ for each pre-activated $z_d^{(l+1)}$ can be found using the function $f(x)$, with layer l close to the input layer.

$$f(x) = \Sigma_{d=1}^{V(l+1)} Rel\_S_d^{(l+1)} = \Sigma_{d=1}^{V(l)} Rel\_S_d^{(l)} = \cdots = \Sigma_{i=1}^{V(1)} \tag{2}$$

The equation mentioned above is back-propagated from the final layers (classifier output $f(x)$) to the input layer x, which contains the pixels of the image. LRP performs backward propagation from the classifier towards the input. While the model predicts in the forward direction, LRP operates in the opposite direction to generate evidence supporting the prediction. The fundamental procedure for LRP is shown clearly in Figure. 4

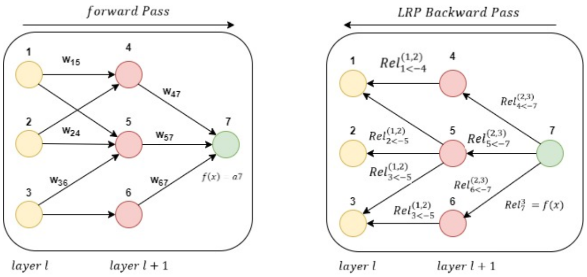


FIGURE 4. Fundamental procedure for Layer-wise relevance propagation, relevance score Rel_j^{l+1}, $j \in \{1, 2...., n^{l+1}\}$ of the j^{th}component for the layer $l + 1$.

G. RANKING EMOTIONS USING LRP RELEVANCE SCORES

In our study, we presented a new approach for evaluating relevance and ranking emotions for accurately predicted images in the test dataset, using Layer-wise Relevance Propagation (LRP) backpropagation. Once the relevance scores $Rel\_S(xi) = \{r_1, r_2, \ldots, r_n\}$ were obtained for each image, we sorted these scores into ten fixed-width intervals (groups) according to their positive values, ranging from [0.0 to 1.0], with each interval covering a range of 0.1. The groups are defined as follows: Group 1: $[0.0, 0.1)$, Group 2: $[0.1, 0.2)$, Group 3: $[0.2, 0.3]$, Group 4: $[0.3, 0.4]$, Group 5: $[0.4, 0.5]$, Group 6: $[0.5, 0.6]$, Group 7: $[0.6, 0.7]$, Group 8: $[0.7, 0.8]$, Group 9: $[0.8, 0.9]$, and Group 10: $[0.9, 1.0]$. This process of using fixed-width interval binning with 10 bins indeed tries to determine the central tendency of the relevance scores in different sections of the interval range [0.0, 1.0].

For each relevance score r_i, we determined its corresponding interval I_j based on its value:

$$I_j = \{r_i : \frac{j-1}{10} \leq r_i < \frac{j}{10}\}.$$

Next, we calculated the mean relevance scores for three groups of intervals. For the initial 5 intervals (Group 1 to 5), we calculated the mean μ_1 as:

$$\mu_1 = \frac{1}{\sum_{i=1}^{5} |I_i|} \sum_{i=1}^{5} \sum_{r \in I_i} r.$$

For the central 5 intervals (Group 4 to 8), the mean μ_2 was calculated as:

$$\mu_2 = \frac{1}{\sum_{i=4}^{8} |I_i|} \sum_{i=4}^{8} \sum_{r \in I_i} r.$$

Finally, for the last 5 intervals (Group 6 to 10), we computed the mean μ_3 as:

$$\mu_3 = \frac{1}{\sum_{i=6}^{10} |I_i|} \sum_{i=6}^{10} \sum_{r \in I_i} r.$$

After calculating the values of μ_1, μ_2, and μ_3, we compared them to determine the emotion rank for each image. If μ_1 was the greatest, the image was tagged with the emotion rating ”Minimal.” If μ_2 had the highest value, the image was given a rank of ”Average.” If μ_3 produced the highest average, the image was labeled as ”Strong”. This new method of analyzing relevance scores and ranking based on means allowed us to precisely evaluate the importance of emotions in every image and place them in suitable emotion rank clusters. By using this approach, we sought to further understand the model's effectiveness and improve the precision of emotional ranking forecasts in our study. Detailed flow of the process is represented in Fig. 5

IV. RESULTS AND DISCUSSION

This section unveils the main discoveries of the study and provides a thorough analysis of the results. The research focused on examining emotions and categorizing them into ranks—Minimal or Average, or Strong.

A. DATASETS USED FOR VALIDATION OF THE MODEL

In this section, we discuss the popular benchmark datasets for the FER problem, which include Extended Cohn-Kanade (CK+) [33], Karolinska Directed Emotional Faces (KDEF) [34], Japanese Female Facial Expression (JAFFE) [35], and AffectNet [36] datasets. These datasets are widely used in our research experiments. Examples of the images in the dataset are displayed in Figure 6. The images represent seven emotional classes: Anger (An), Contempt (Co), Disgust (Di), Fear (Fe), Happy (Ha), Surprise (Su), and Sadness (Sa). A brief description of the datasets is as follows:

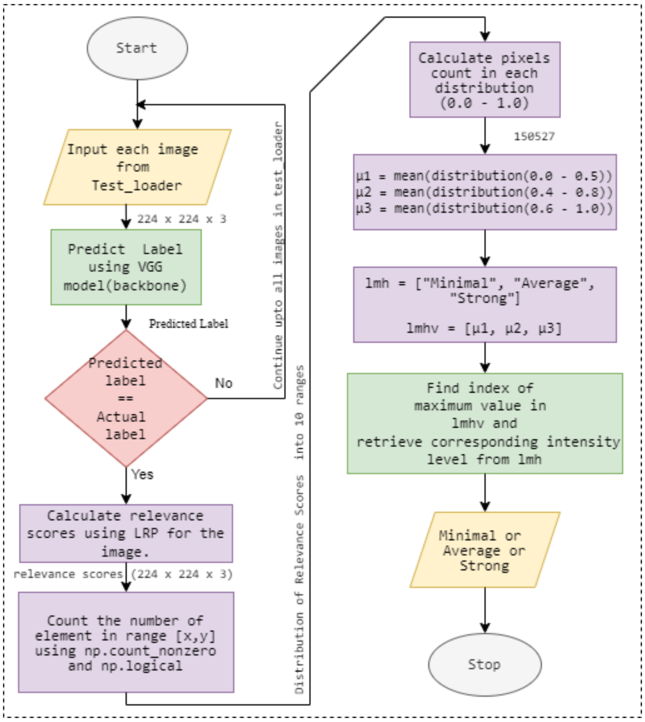

Start
Input each image from Test_loader
224 x 224 x 3
Predict Label using VGG model(backbone)
Predicted Label
Predicted label == Actual label
No
Continue upto all images in test_loader
Yes
Calculate relevance scores using LRP for the image.
relevance scores (224 x 224 x 3)
Count the number of element in range [x,y] using np.count_nonzero and np.logical
Distribution of Relevance Scores into 10 ranges
Calculate pixels count in each distribution (0.0 - 1.0)
150527
µ1 = mean(distribution(0.0 - 0.5))
µ2 = mean(distribution(0.4 - 0.8))
µ3 = mean(distribution(0.6 - 1.0))
lmh = ["Minimal", "Average", "Strong"]
lmhv = [µ1, µ2, µ3]
Find index of maximum value in lmhv and retrieve corresponding intensity level from lmh
Minimal or Average or Strong
Stop

FIGURE 5. Detailed process of assigning ranks using LRP.

1) CKPlus

The CKPlus dataset serves as an enhancement to the Cohn-Kanada dataset, featuring images with a resolution of 48 by 48 pixels. In total, there are 981 images available. Widely recognized as one of the most extensively used laboratory-controlled facial expression classification databases, the CK+ dataset is a staple for numerous facial expression classification algorithms. Notably, the dataset initially lacked the emotion of contempt, but it has been added in the latest version.

2) KDEF

The KDEF emotion dataset is well-known and easily accessible for studying facial expressions and recognizing emotions. This dataset, created by the Karolinska Institute in Sweden, includes 4900 high-resolution pictures with color, showcasing 70 unique people (35 women and 35 men) who are between the ages of 20 and 30. These people display seven different facial expressions, such as joyous, sorrowful, enraged, scared, shocked, repulsed, and indifferent. Every individual is captured from 7 distinct angles, including left and right profiles, and front-facing shots featuring different facial expressions. The KDEF dataset's versatility is an im-

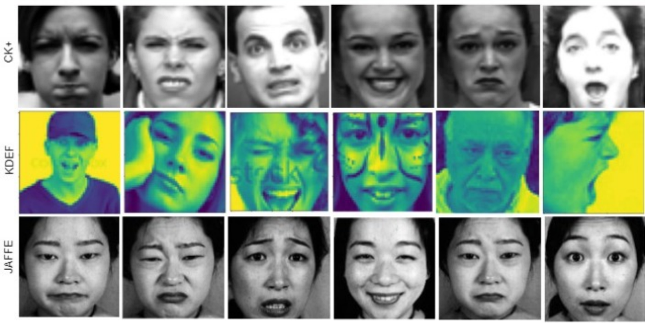


FIGURE 6. Sample images from datasets

portant resource for a wide range of emotion analysis and facial recognition tasks. Researchers and professionals often use the KDEF dataset for comparing and testing algorithms, due to its diverse range of facial expressions captured in controlled settings, making it easier to conduct thorough studies on emotions.

3) JAFFE

The JAFFE emotion dataset is significant in the field of analyzing facial expressions and recognizing emotions. Produced by Japanese scientists, this dataset includes 213 black and white pictures showcasing 10 Japanese female models displaying seven different facial expressions. Every model displays a neutral face, as well as six basic emotions: Anger(An), Contempt (Co), Disgust(Di), Fear(Fe), Surprise(Su), and Sadness(Sa). The pictures were captured in controlled light settings to maintain uniformity. Providing a thorough selection of facial expressions from various people, the JAFFE dataset is a useful tool for emotion-related research. Researchers and developers often use it to assess and teach facial expression recognition systems. Table. 2. shows the distribution of samples in the JAFFE, CK+, and KDEF datasets.

TABLE 2. Sample distribution of different datasets across various emotions. An:Anger, Co:Contempt, Di:Disgust, Fe:Fear, Ha:Happy, Sa:Sad, Su:Surprise.

Dataset	An	Co	Di	Fe	Ha	Sa	Su	Total
JAFFE	30	30	30	30	30	30	30	180
CK+	135	54	177	75	207	84	249	981
KDEF	70	70	70	70	70	70	70	490

4) AffectNet

AffectNet is a large-scale facial emotion dataset containing images collected from the web through search engines. It is specifically designed to support the development and evaluation of facial expression recognition algorithms in conditions that closely resemble real-world scenarios. Since the images in AffectNet are taken in uncontrolled settings, the dataset is especially useful for creating reliable algorithms that can manage the complexity and unpredictability of real-world situations. Sample images of the Affectnet dataset are represented in Fig. 7 where uncontrolled settings are clearly visible. Variations in illumination, a range of backgrounds, various facial expressions, and partial occlusions are well-represented and make it difficult for models to function consistently outside of lab environments. The number of samples distributed for each emotion class of this dataset is shown in Fig. 8. The image in the second row of Fig. 8 showing a neutral expression has the face partially covered by a hand. Additionally, the images depicting happiness, fear, and surprise include occlusions caused by glasses.

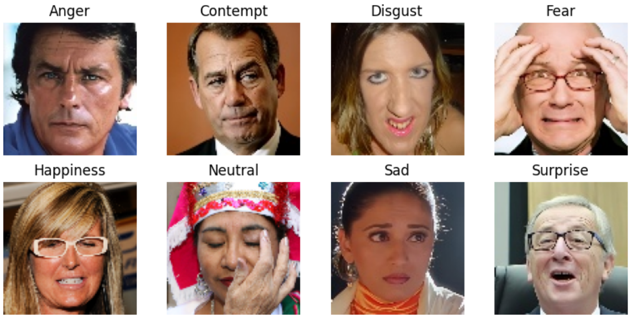


FIGURE 7. Sample images of AffectNet Dataset.

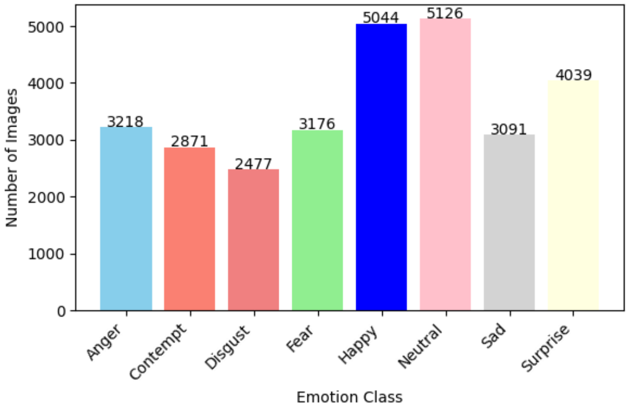


FIGURE 8. Sample distribution for each emotion of AffectNet Dataset.

B. GENERATION OF IMAGE HEATMAPS

A heatmap is a graphical representation of the relative importance of different parts of an image to a specific task or decision made by a deep learning model. We employed the Layerwise Relevance Propagation (LRP) method to generate heatmaps, providing insights into the decision-making process of the VGG-16 model. In heatmap visualizations, the importance of each pixel is shown by its color. A color scale is typically used to depict the relevance scores: cool colors (such as blue or green) indicate low significance, while warm colors (such as yellow, orange or red) indicate high relevance. In this study, blue and red were utilized as the color scale. For instance, in the CKPlus dataset, Fig. 9 shows two images and their respective heatmaps representing the emotions of

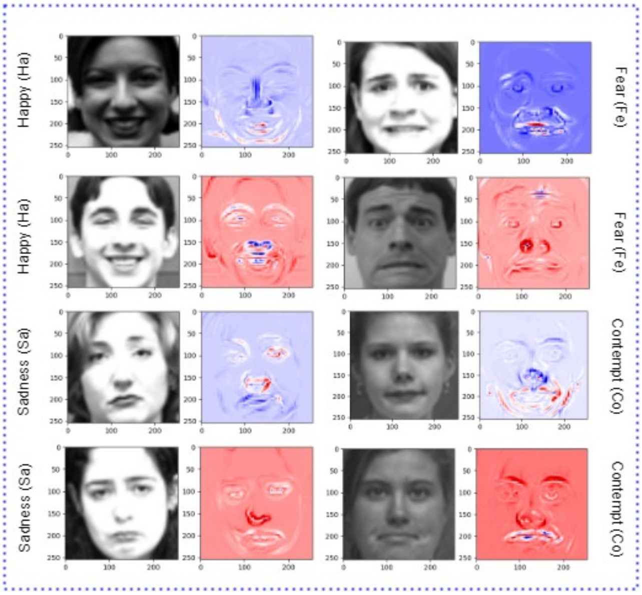


FIGURE 9. Visualization of Heat Maps for few random images.

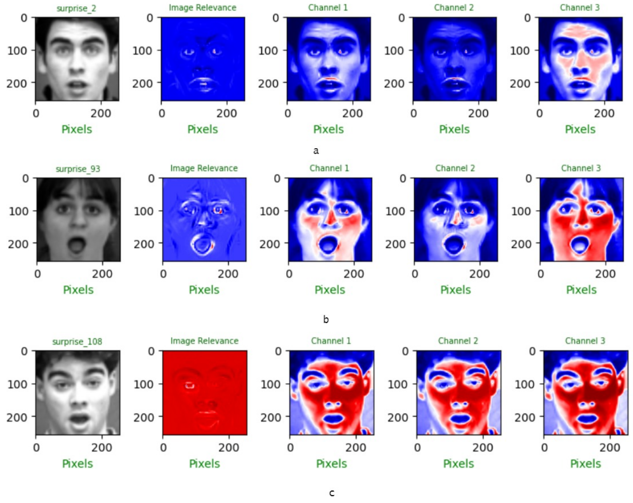


FIGURE 10. Image relevance and heatmaps for three channels (RGB) from the CKPlus dataset image. a. Image relevance for Minimal rank. b. Image relevance for Average rank. c. Image relevance for Strong rank.

happiness(Ha) sadness (Sa), fear (Fe), and contempt (Co). In these heatmaps, blue signifies low relevance, whereas red indicates high relevance to the given image. Heatmaps make it easy to identify regions of an image that are highly relevant and those that are not. We use cool colors to represent areas that are not very important to the model's outcome and warm colors to depict regions considered highly relevant by the model's decision-making process. The color gradient from cool to warm highlights the areas that matter most for the model's prediction, providing crucial information about the model's classification features. For example, the heatmap of the contempt (Co) images in the last row of Fig. 9 is highly relevant, as indicated by the dominance of the red color. By examining these heatmaps, we can pinpoint exact regions (such as areas around the mouth, nose, eyes, etc.) that significantly impact the model's output. This information helps in understanding the model's behavior, identifying its biases and shortcomings, and gaining insight into the assumptions underpinning its predictions. Heatmaps are an essential part of our research on understanding the VGG-16 model with Layerwise Relevance Propagation. They enable us to validate and comprehend the generalizability and robustness of deep learning models.

Heatmaps are an essential part of our research on understanding the VGG-16 model with Layerwise Relevance Propagation because they allow us to validate and understand the generalizability and robustness of deep learning models.

C. RANKINGS OF EMOTIONS WITH LRP

In order to determine the intensity of someone's feelings, ranking them is a crucial part of human face emotion recognition. To rank emotions, a numerical score or intensity level is assigned to each identified emotion, indicating the perceived

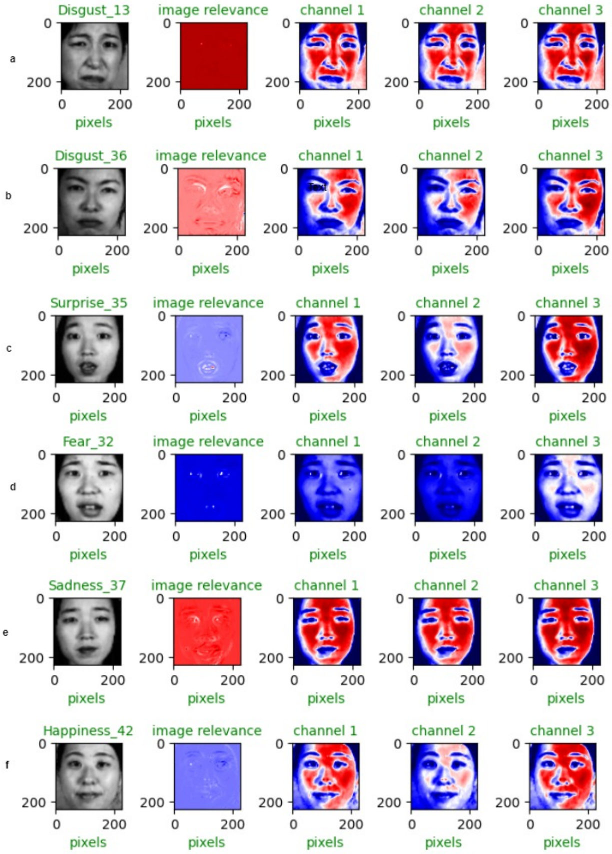
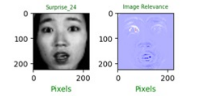
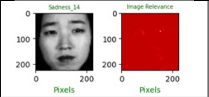
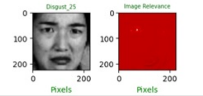
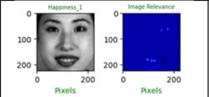
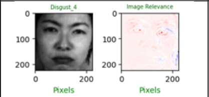
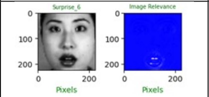
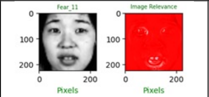
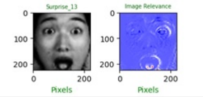


FIGURE 11. Image relevance scores and heatmaps for the RGB channels of the JAFFE dataset. a. Disgust image with Strong intensity rank. b. Disgust face with Average intensity rank. c. Surprise image with Minimal intensity rank. d. Fear image with Minimal intensity rank. e. Sadness image with Strong intensity rank. f. Happiness image with Minimal intensity rank.

TABLE 3. Table of image relevance scores and assigned ranks (Minimal, Average, Strong) for emotions in the JAFFE dataset.

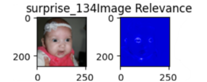
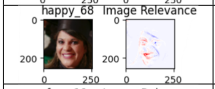
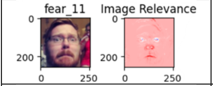
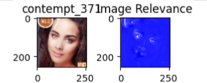
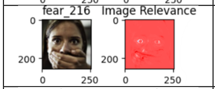
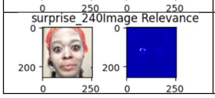
Image	Group										Mean			Rank
Image Relevance	0 - 1	1 - 2	2 - 3	3 - 4	4 - 5	5 - 6	6 - 7	7 - 8	8 - 9	9 - 10	μ_1	μ_2	μ_3	
	1	9	38	728	149542	174	23	8	2	2	150318	150475	209	Average
	0	0	0	3	1	7	18	139	150155	204	4	168	150523	Strong
	1	1	2	2	6	9	18	82	147342	3064	12	117	150515	Strong
	341	148454	1455	196	51	16	6	6	1	1	150497	275	30	Minimal
	1	3	11	283	147445	2720	47	12	2	3	147743	150507	2784	Average
	1	42	149906	473	62	21	10	5	5	2	150484	571	43	Minimal
	1	0	6	14	39	209	3959	145999	292	8	60	150220	150467	Strong
	3	32	325	146183	3601	295	72	11	2	3	150144	150162	383	Average

strength or intensity of that emotion in a specific facial expression. We use a scale to rank emotions from 1 to 3, labeled as "MINIMAL," "AVERAGE," and "STRONG." The input image is a 224 x 224 colored image with dimensions 224 by 224 by 3. Some pixels have a favorable impact, while others have a negative one. We consider the overall number of pixels that contributed positively. During our experimentation, we assigned intensity ranks based on Layerwise Relevance Propagation (LRP) using the JAFFE, CKPlus, KDEF, and AffectNet datasets. The intensity ranks "Minimal," "Average," and "Strong" based on relevance scores are represented in Tables 3 and 4 for the JAFFE and AffectNet datasets, respectively. These tables provide a comprehensive view of the results from both controlled (JAFFE) and unconstrained (AffectNet) environments.

D. PILOT EXPERIMENTATION FOR GROUND TRUTH

This research aims to assess the strength of facial expressions by assigning appropriate labels. Our study introduces a new method of annotating images of human facial emotions by leveraging input from multiple annotators to enhance the strength and reliability of emotion rankings. We involved a total of 10 postgraduate volunteers, aged between 25 and 30, with 7 males and 3 females, in the annotation task. Each reviewer independently assessed images depicting different emotions, shown sequentially. For instance, each of the 10 annotators evaluated an image from the "Happy" category and classified the observed emotional intensity as "Minimal," "Average," or "Strong" after thoroughly analyzing the facial expressions. We employed a majority voting method to establish the emotion ranking for each image, selecting the rank with the most agreement among the 10 annotators as the final emotion ranking. For example, if all 10 annotators agreed on ranking a specific "Happy" image as "Minimal," then the emotion rank for that image would be labeled as "Minimal Happy." Similarly, the rankings for "Average Happy" and "Strong Happy" were determined by the majority vote. To ensure comprehensive coverage, each annotator assessed images from every emotion class. Table. 5 provides a breakdown of how the images are divided among annotators for the CKPlus, KDEF, JAFFE, and AffectNet datasets. All images from each class were assigned to every annotator for these datasets. By using this collaborative and iterative method of emotion annotation, we aimed to reduce personal biases and achieve higher precision and reliability in emotion rankings. This new annotation approach provided us with valuable insights into human perception of facial emotions, forming a robust foundation for training and testing our emotion ranking model. Table. 5 displays summary sheets containing the labels assigned by the annotators for the KDEF dataset. This experimental process was applied to each of the four datasets. Due to the unconstrained and

TABLE 4. Table of image relevance scores and assigned ranks (Minimal, Average, Strong) for emotions in the AffectNet dataset.

Image	Group										Mean			Rank
Image Relevance	0 - 1	1 - 2	2 - 3	3 - 4	4 - 5	5 - 6	6 - 7	7 - 8	8 - 9	9 - 10	μ_1	μ_2	μ_3	
	3	0	7	11	69	366	189996	4532	83	7	90	194974	194984	Strong
	43	193132	1756	103	25	8	4	1	1	1	195059	141	15	Minimal
	2	5	15	796	193695	497	54	6	1	3	194513	195048	561	Average
	2	5	17	26	129	476	192818	1512	73	16	179	194961	19489	Average
	2	87	181651	12657	548	93	19	12	3	2	194945	13329	129	Minimal
	1	5	13	33	96	369	190975	3500	67	15	148	194973	194926	Average
	1	3	7	8	28	73	330	187173	7354	97	47	187612	195027	Strong
	315	194212	388	87	44	18	2	4	2	2	195046	55	28	Average

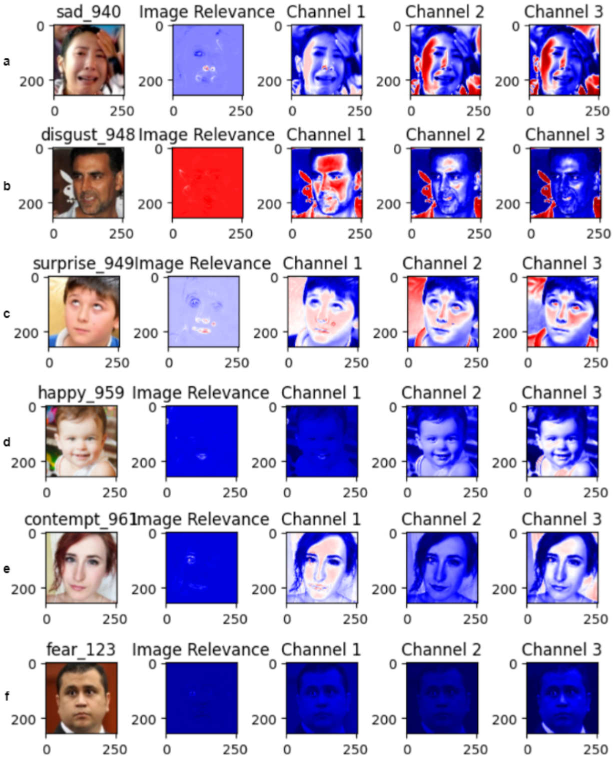


FIGURE 12. Image relevance scores and heatmaps for the RGB channels of the AffectNet dataset. a. Sad image with Minimal intensity rank. b. Disgust image with Strong intensity rank. c. Surprise image with Average intensity rank. d. Happy image with Minimal intensity rank. e. Contempt image with Minimal intensity rank. f. Fear image with Minimal intensity rank.

large size of the AffectNet dataset, pilot experimentation for these images was notably more challenging and time-consuming compared to the lab-controlled datasets CKPlus, JAFFE, and KDEF. Nevertheless, we successfully overcame these challenges through our rigorous annotation process.

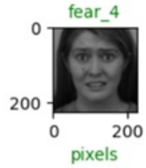
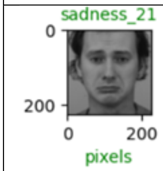
E. VALIDATION OF RANKS THROUGH RELEVANCE SCORES ALONGSIDE PILOT EXPERIMENTATION RANKS

The ranking labels predicted using image relevance scores are cross-validated with the ground truth labels assigned by the 10 annotators who participated in the annotation process. The predicted labels, generated using Layer-wise Relevance Propagation (LRP), are validated against the volunteer-annotated labels for the CK+ dataset through a confusion matrix analysis, as shown in Fig. 16. Similarly, the validation for the JAFFE dataset is represented in Fig. 15. It is important to note that the JAFFE dataset is minimal in size, with a very small number of samples in the test dataset split. Fig. 17 illustrates the cross-validated intensity ranks for the AffectNet dataset.

F. PERFORMANCE METRICS

In this subsection, we discuss the performance metrics of our emotion classification model, focusing on accuracy, precision, recall, and F1-score for each emotion class and intensity rank. The performance of our model has been evaluated using two datasets: KDEF, JAFFE, CKPlus and AffectNet. Tables 6, 7, 8 and 9 summarize these metrics for each emotion class and rank. Table 6 presents the performance metrics for

TABLE 5. Annotators Summary Sheet of few samples of KDEF dataset for labeling Low, Medium, and High Rank for the images. Ranks: Min - Minimal, Avg - Average, Str - Strong. Annotators: $A_1, A_2, \ldots, A_{10}$.

Image	Annotators										Label Count			Final Label
	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8	A_9	A_{10}	Min	Avg	Str	
	Min	Min	Avg	Min	Min	Min	Min	Min	Avg	Min	8	2	0	Min
	Min	Avg	Avg	Str	Min	Avg	Avg	Avg	Str	Avg	2	6	1	Avg
	Str	Avg	Avg	Str	Str	Avg	Str	Str	Str	Str	0	3	7	Str
	Min	Min	Min	Avg	Min	Min	Min	Min	Min	Min	9	1	0	Min
	Str	Avg	Avg	Avg	Avg	Avg	Avg	Str	Str	Avg	0	7	3	Avg
	Min	Min	Min	Avg	Min	Avg	Min	Min	Avg	Min	7	3	0	Min

the KDEF dataset. Our model demonstrates robust accuracy across all emotion classes and intensity ranks, with accuracy values ranging from 0.91 to 0.97, and an overall average accuracy of 0.93. These high accuracy values indicate that the model is consistently reliable in predicting the correct emotion classes. For the emotion 'Anger,' the model achieves an accuracy of 0.94 across all intensity ranks, demonstrating high reliability in detecting anger. Similarly, the accuracy for 'Disgust' is consistent at 0.91, indicating a strong performance. The model achieves an accuracy of 0.92 for 'Fear,' showing it can effectively recognize this emotion. The highest accuracy of 0.97 is observed for the 'Happy' emotion, indicating the model's exceptional performance in detecting happiness. The accuracy for 'Sadness' is 0.91, reflecting solid performance, while for 'Surprise,' the model achieves an accuracy of 0.93, demonstrating good performance in recognizing surprise.

Table 7 presents the performance metrics of our emotion classification model on the JAFFE dataset, showcasing the accuracy, precision, recall, and F1-score for different emotion classes at varying intensities (Minimal, Average, Strong). The model demonstrated exceptional performance for the emotion class 'Fear', achieving perfect scores (1.00) across all metrics and intensities. Similarly, for 'Happy', the model maintained perfect accuracy and F1-scores across all intensity levels, although a slight drop in precision (0.67) and recall (0.67) was noted at the Strong intensity. 'Sadness' and 'Disgust' exhibited a slight decline in performance with an average precision of 0.50 and 0.67, respectively, particularly at the Strong intensity. For 'Surprise' and 'Anger', the model again showed perfect accuracy across all intensities, with a minor reduction in recall (0.67) for Strong 'Surprise'. The

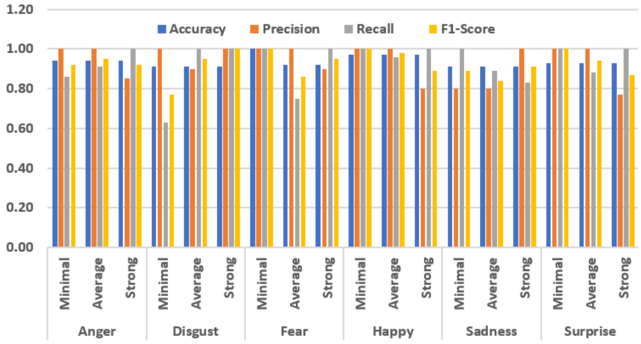


FIGURE 13. Accuracy, Precision, Recall and F1-Score for CKPlus dataset.

overall average performance across all emotion classes and intensities was high, with an accuracy of 0.96, precision of 0.89, recall of 0.83, and an F1-score of 0.83, indicating a robust and reliable classification performance by the model. Table 8 presents the performance metrics for the CKPlus dataset. The accuracy values for this dataset are also impressive, with an overall average accuracy of 0.93. The model shows high accuracy of 0.94 across all intensity ranks for ’Anger,’ similar to the KDEF dataset. The accuracy for ’Disgust’ is slightly lower at 0.91 but still reflects a strong performance. The accuracy for ’Fear’ is 0.92, consistent with the performance on the KDEF dataset. For the ’Happy’ emotion, the accuracy is 0.97, indicating excellent detection capability. The accuracy for ’Sadness’ is 0.91, demonstrating the model’s reliability, and for ’Surprise,’ the model achieves an accuracy of 0.93, similar to its performance on the KDEF dataset.

Table 9 illustrates the performance metrics of our emotion classification model on the AffectNet dataset, detailing the accuracy, precision, recall, and F1-score for various emotion classes across different intensity levels (Minimal, Average, Strong). For ’Anger’, the model achieved high performance, with accuracy and F1-scores consistently around 0.95 across all intensities. ’Disgust’ showed a slight variation, with accuracy and F1-scores ranging from 0.88 to 0.91. The ’Fear’ emotion class displayed a consistent performance with an overall accuracy of 0.89 and a slightly lower recall at 0.85. The model performed exceptionally well in classifying ’Happy’ emotions, maintaining high precision and recall around 0.92 to 0.93. For ’Sadness’, a slight dip in performance was observed, especially at the Average intensity, with accuracy at 0.88 and recall at 0.86. The ’Surprise’ class also showed strong results, with metrics consistently around 0.90. Overall, the model achieved an average performance across all emotions and intensities with an accuracy of 0.92, precision of 0.91, recall of 0.91, and an F1-score of 0.91, indicating a robust and reliable classification capability on the AffectNet dataset. These metrics highlight the model’s strengths in accurately classifying emotions, which consistently shows the highest accuracy.

TABLE 6. Performance Metrics of Emotion Classification by Emotion Class and Rank for the KDEF Dataset.

Emotion	**Class**	**Accuracy**	**Precision**	**Recall**	**F1-Score**
Anger	Minimal	0.94	1.00	0.85	0.92
	Average	0.94	1.00	0.90	0.95
	Strong	0.91	0.85	1.00	0.92
Disgust	Minimal	0.91	1.00	0.60	0.77
	Average	0.94	1.00	1.00	1.00
	Strong	0.91	1.00	1.00	1.00
Fear	Minimal	0.91	1.00	1.00	1.00
	Average	0.92	1.00	0.80	0.89
	Strong	0.91	0.85	1.00	0.92
Happy	Minimal	0.97	1.00	1.00	1.00
	Average	0.97	1.00	1.00	1.00
	Strong	0.97	0.87	1.00	0.93
Sadness	Minimal	0.91	1.00	0.80	0.89
	Average	0.91	1.00	0.88	0.93
	Strong	0.91	0.80	0.80	0.80
Surprise	Minimal	0.93	1.00	1.00	1.00
	Average	0.93	1.00	1.00	1.00
	Strong	0.93	0.77	1.00	0.87
Average		**0.93**	**0.96**	**0.89**	**0.91**

TABLE 7. Performance Metrics of Emotion Classification by Emotion Class and Rank for JAFFE Dataset.

Emotion	Class	Accuracy	Precision	Recall	F1-Score
Anger	Minimal	1.00	1.00	1.00	1.00
	Average	1.00	1.00	1.00	1.00
	Strong	1.00	1.00	1.00	1.00
Fear	Minimal	1.00	1.00	1.00	1.00
	Average	1.00	1.00	1.00	1.00
	Strong	1.00	0.00	0.00	0.00
Disgust	Minimal	1.00	1.00	1.00	1.00
	Average	0.80	1.00	0.50	0.67
	Strong	1.00	1.00	1.00	1.00
Happy	Minimal	1.00	1.00	1.00	1.00
	Average	1.00	1.00	1.00	1.00
	Strong	0.80	1.00	0.50	0.67
Sadness	Minimal	1.00	1.00	1.00	1.00
	Average	0.86	0.50	0.50	0.50
	Strong	0.80	0.50	0.50	0.50
Surprise	Minimal	1.00	1.00	0.67	0.80
	Average	1.00	1.00	1.00	1.00
	Strong	1.00	1.00	1.00	1.00
Average		**0.96**	**0.89**	**0.83**	**0.83**

G. CONFUSION MATRIX ANALYSIS

Confusion matrices were created to compare the model’s predicted intensity rank labels (Minimal, Average, and Strong) for each emotion against the labels assigned by voluntary annotators during Pilot Experimentation. For each emotion, a confusion matrix summarizes how often the predicted labels match the annotators’ labels, displaying the number of correct predictions along the diagonal and the misclassifications in the off-diagonal cells. This allows for a detailed analysis of the model’s performance in accurately predicting the intensity of different emotions. Confusion matrix examinations were performed on both the JAFFE, CKPlus and AffectNet datasets, and the outcomes are depicted in Fig(s). 15, 16, 17. The JAFFE dataset, comprising merely 180 samples, is relatively small, leading to only 5 samples being present in the test dataset. In the confusion matrix pertaining to the Anger emotion, depicted in Fig. 15 it is evident that two

TABLE 8. Performance Metrics of Emotion Classification by Emotion Class and Rank for CKPlus Dataset.

Emotion	Class	Accuracy	Precision	Recall	F1-Score
Anger	Minimal	0.97	1.00	0.85	0.92
	Average	0.97	1.00	0.91	0.95
	Strong	0.95	0.85	1.00	0.92
Disgust	Minimal	0.94	1.00	1.00	0.97
	Average	0.95	1.00	1.00	0.97
	Strong	0.95	1.00	1.00	0.97
Fear	Minimal	0.96	1.00	0.80	0.89
	Average	0.97	1.00	0.75	0.86
	Strong	0.96	1.00	0.89	0.94
Happy	Minimal	0.97	1.00	1.00	1.00
	Average	0.97	1.00	1.00	1.00
	Strong	0.97	0.87	1.00	0.93
Sadness	Minimal	0.96	1.00	0.80	0.89
	Average	0.95	1.00	0.88	0.93
	Strong	0.96	1.00	0.93	0.97
Surprise	Minimal	0.96	1.00	1.00	1.00
	Average	0.96	1.00	0.89	0.94
	Strong	0.96	0.77	1.00	0.87
Average		**0.9578**	**0.96**	**0.92**	**0.93**

TABLE 9. Performance Metrics of Emotion Classification by Class and Rank for AffectNet Dataset.

Emotion	Class	Accuracy	Precision	Recall	F1-Score
Anger	Minimal	0.96	0.94	0.96	0.95
	Average	0.96	0.95	0.95	0.95
	Strong	0.96	0.94	0.96	0.95
Disgust	Minimal	0.93	0.91	0.91	0.91
	Average	0.94	0.91	0.91	0.91
	Strong	0.94	0.91	0.91	0.91
Fear	Minimal	0.94	0.89	0.88	0.89
	Average	0.95	0.89	0.89	0.89
	Strong	0.95	0.89	0.89	0.89
Happy	Minimal	0.96	0.91	0.91	0.91
	Average	0.96	0.92	0.90	0.91
	Strong	0.96	0.92	0.90	0.91
Sadness	Minimal	0.94	0.88	0.88	0.89
	Average	0.94	0.86	0.86	0.86
	Strong	0.94	0.88	0.88	0.89
Surprise	Minimal	0.94	0.90	0.90	0.90
	Average	0.95	0.90	0.89	0.90
	Strong	0.95	0.90	0.89	0.90
Average		**0.94**	**0.91**	**0.90**	**0.90**

Average-Anger images were incorrectly classified as Strong-Anger. Similarly, the CKPlus dataset has a total number of samples of 981. Out of which 20% is considered as a test dataset split. Using these samples a confusion matrix plot is evaluated which is shown in Fig. 16. The AffectNet dataset is a dataset connected to real-time scenarios. We have also considered this dataset for experimentation. The correctly predicted samples from the test dataset split were selected for analysis, excluding the Neutral emotion due to its negligible importance in intensity ranking. The remaining samples were assigned intensity rank labels of Minimal, Average, and Strong which is shown in Fig. 17.

H. RESULTS COMPARISON WITH THE OTHER EXISTING WORKS

In this section, to contextualize the performance of our proposed model, we compare it with several worlds in the field of FER. Table. 10 presents the summary of the comparisons, techniques used, the intensity ranks, datasets used, accuracy achieved and noted drawback or advantage. The research work by Siti Khauruni et al. [24] has utilized Hidden Markov Model (HMM) to predict the emotion intensity into 3 categories namely Neutral, Onset and Apex states. They used CKPlus dataset, achieving an accuracy of 82.3%. The significant challenge of this approach is the necessity of geometric feature extraction which is computationally high which may not be generalize well to varied datasets. Kai-Tai et al. [37] employed texture features with backpropogation on an unspecified dataset, achieving a high accuracy of 95.6%. Despite of impressive accuracy this technique is identified that it is not suitable for real-time applications, limiting its practical use.

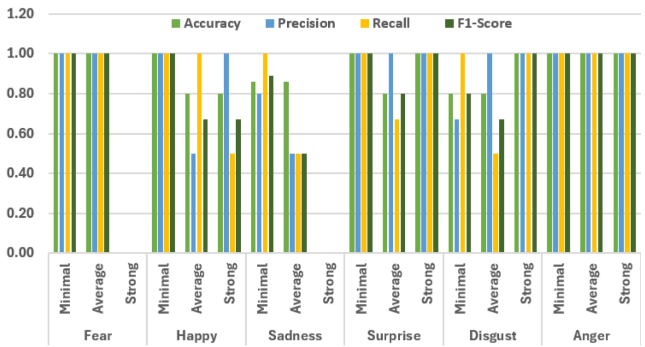


FIGURE 14. Accuracy, Precision, Recall and F1-Score for JAFFE dataset.

ML-DCNNet [35], which leveraged deep convolutional nerual networks, evaluated the emotion intensities ranging from Neutral to Apex across the CKPlus and JAFFE datasets. This method has achieved high accuracy rates of 99.15% and 93.43% respectively. However, this model's performance is validated on lab-controlled datasets, which are not reflecting real-world scenarios. Kuang-Yu [38] utilized descriptor scattering transform with intensity parameters grouped as Low, Medium and High. This method yielded an accuracy of 68.2% revealing its unsuitability for real-time applications due to its lower performance. Rohit Pal et. al [39] proposed a clustering technique combined with backpropagation on the CKPlus dataset achieving an accuracy of 98.00%. This technique also has a challenge of real-time applicability. Jane Reilly et al. [40] adopted an image-based approach with Support Vector Machine (SVM) classifiers, assessing emotions across Low, Medium and High intensities. They worked with a very small dataset of 68 images from the CKPlus dataset, achieving an accuracy of 89% which raised concerns about the generalizability of their results.

Our proposed model employs Layer-wise Relevance Propagation (LRP) of the XAI framework to assign emotion intensity ranks of Minimal, Average and Strong. Evaluated across multiple datasets: CKPlus, JAFFE, KDEF and AffectNet. The proposed model demonstrates robust performance with accuracies of 95.79%, 90.33%, 93.78% and 93.89%

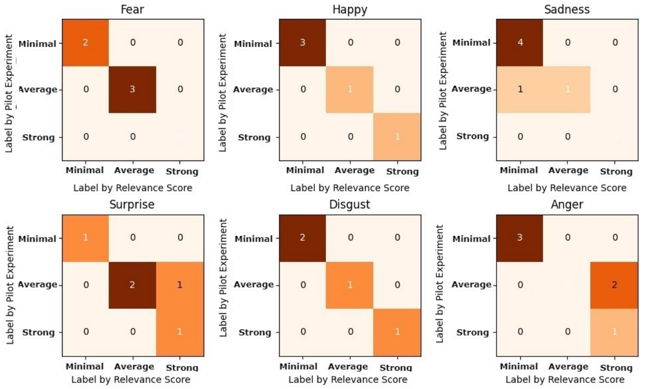


FIGURE 15. Confusion matrix for the annotated labels and predicted labels by image relevance for the JAFFE dataset.

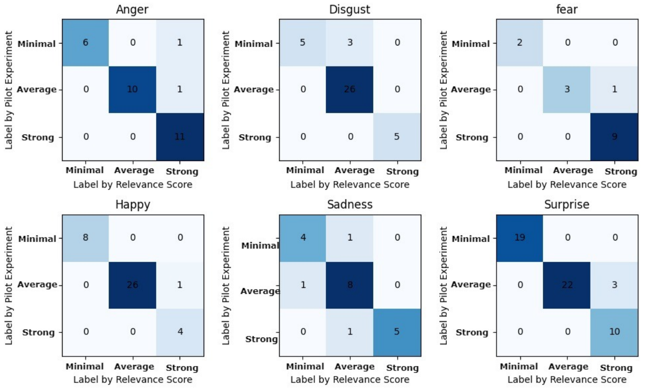


FIGURE 16. Confusion Matrix for the annotated labels and predicted labels by image relevance for CKPlus Dataset

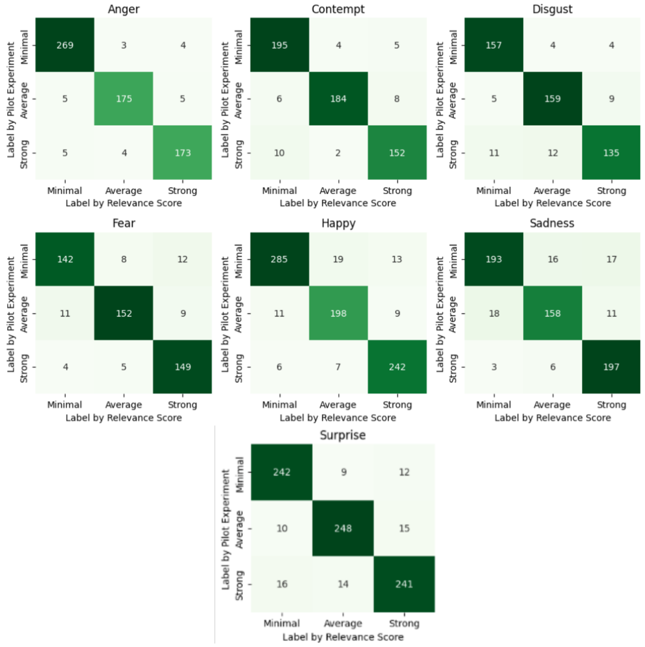


FIGURE 17. Confusion Matrix for the annotated labels and predicted labels by image relevance for AffectNet Dataset.

respectively. This approach not only achieves high accuracy but also provides the advantage of being applicable to both lab-controlled and real-time environments.

V. CONCLUSION, LIMITATION AND POTENTIAL FUTURE SCOPE

In conclusion, our study highlights the effectiveness of using the VGG-16 pre-trained model with transfer learning for facial emotion recognition. This approach yields remarkable accuracies across diverse datasets, demonstrating the model's adaptability. The utilization of Layerwise Relevance Propagation (LRP) as an eXplainable Artificial Intelligence (XAI) technique enhances our understanding of the model's decision-making process, providing valuable insights into emotion ranking based on relevance scores. Our novel approach for predicting intensity ranks of emotion, combined with emotion labels, not only improves interpretability but also fosters trust in deep learning(DL) models for the analysis of emotion. The successful validation of assigned ranks by independent annotators adds credibility to our findings, contributing to the domain of FER. Our research demonstrates the capacity of eXplainable Artificial Intelligence (XAI) to enhance the resilience and transparency of emotion recognition systems. Our research is limited to basic human facial emotions and not with emotions like Pride, Guilt, Shame, Jealousy, Love, Nostalgia, Embarrassment, Relief, Gratitude, Admiration. Here these emotions are mixed emotions i.e., Pride is a combination of happiness and surprise. Looking forward, future research in facial emotion recognition could explore alternative XAI techniques beyond LRP, offering diverse perspectives on model interpretability. Investigating bigger and varied datasets, including real-life situations, can enhance the overall effectiveness and usability of facial emotion recognition models. Explore more emotions rather than basic emotions which help the domain of FER. Researching new methods for ranking emotions and comparing them to various existing methods would enhance our understanding of this field. Close attention should be given to the ethical implications of utilizing facial emotion data and XAI in order to guarantee the responsible and fair implementation of emotion recognition systems in a variety of applications. To summarize, our study provides a solid basis for future efforts to improve facial emotion detection and enhance XAI methods in order to develop more clear and dependable emotion analysis models.

REFERENCES

[1] Sudheer Babu Punuri, Sanjay Kumar Kuanar, Manjur Kolhar, Tusar Kanti Mishra, Abdalla Alameen, Hitesh Mohapatra, and Soumya Ranjan Mishra. Efficient net-xgboost: an implementation for facial emotion recognition using transfer learning. Mathematics, 11(3):776, 2023.

[2] Yacine Yaddaden, Abdenour Bouzouane, Mehdi Adda, and Bruno Bouchard. A new approach of facial expression recognition for ambient assisted living. In proceedings of the 9th ACM international conference on pervasive technologies related to assistive environments, pages 1–8, 2016.

[3] Caifeng Shan, Shaogang Gong, and Peter W McOwan. Facial expression recognition based on local binary patterns: A comprehensive study. Image and vision Computing, 27(6):803–816, 2009.

TABLE 10. Model comparison with other existing works

Work	Technique Used	Intensity Parameters	Dataset	Accuracy	Drawback Advantages
Siti Khairuni et. al [24]	Intensity using HMM	Neutral, Onset, Apex	CKPlus	82.30%	Extraction of Geometric features.
Kai-Tai et al. [37]	Texture features with back propagation	Continous Value Representation	-	95.60%	Not suitable for real-time applications.
ML-DCNNet [35]	Deep Convolutional Neural Network	Neutral - 1, Onset- 2, Offset - 3, Apex - 4	CKPlus JAFFE	99.14% 93.43%	Lab-Controlled Datasets
Kuang-Yu [38]	Descriptor Scattering Transform	Low, Medium High	Single	68.20%	Unsuitable for real-time applications.
Rohit Pal et. al [39]	Clustering of k-means along back-propagation.	-	CKPlus	98.00%	Not suited with real-time applications.
Jane Reilly et. al [40]	Image based approach and Support Vector Machine (SVM)	Low, Medium and High	68 images of 10 subjects from CKPlus	89.00%	Very small dataset.
Proposed Model (Ours)	Layer-wise Relevance Score of XAI	Minimal, Average and Strong	CKPlus JAFFE KDEF	95.79% 96.33% 93.78%	Lab Controlled Dataset
			AffectNet	93.89%	Real-time Applications

[4] Christoph Bartneck and Michael J Lyons. Hci and the face: Towards an art of the soluble. In Human-Computer Interaction. Interaction Design and Usability: 12th International Conference, HCI International 2007, Beijing, China, July 22-27, 2007, Proceedings, Part I 12, pages 20–29. Springer, 2007.

[5] Carlos Orrite, Andrés Ganán, and Grégory Rogez. Hog-based decision tree for facial expression classification. In Pattern Recognition and Image Analysis: 4th Iberian Conference, IbPRIA 2009 Póvoa de Varzim, Portugal, June 10-12, 2009 Proceedings 4, pages 176–183. Springer, 2009.

[6] Magali Batty and Margot J. Taylor. Early processing of the six basic facial emotional expressions. Cognitive Brain Research, 17(3):613–620, 2003.

[7] Asifullah Khan, Anabia Sohail, Umme Zahoora, and Aqsa Saeed Qureshi. A survey of the recent architectures of deep convolutional neural networks. Artificial intelligence review, 53:5455–5516, 2020.

[8] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? Advances in neural information processing systems, 27, 2014.

[9] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. Information fusion, 58:82–115, 2020.

[10] Filip Karlo Došilović, Mario Brčić, and Nikica Hlupić. Explainable artificial intelligence: A survey. In 2018 41st International convention on information and communication technology, electronics and microelectronics (MIPRO), pages 0210–0215. IEEE, 2018.

[11] Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman. Building machines that learn and think like people. Behavioral and brain sciences, 40:e253, 2017.

[12] Erico Tjoa and Cuntai Guan. A survey on explainable artificial intelligence (xai): Toward medical xai. IEEE transactions on neural networks and learning systems, 32(11):4793–4813, 2020.

[13] Xiao Bai, Changsheng Xu, Longzhi Yang, Feiyue Liu, and Huan Zhang. Explainable deep learning for efficient and robust pattern recognition: A survey of recent developments. Pattern Recognition, 120:108102, 2021.

[14] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature machine intelligence, 1(5):206–215, 2019.

[15] Zachary C. Lipton. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. Queue, 16(3):31–57, 2018.

[16] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PloS one, 10(7):e0130140, 2015.

[17] Brian Kenji Iwana, Ryohei Kuroki, and Seiichi Uchida. Explaining convolutional neural networks using softmax gradient layer-wise relevance propagation. In 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), pages 4176–4185. IEEE, 2019.

[18] Aravindh Mahendran and Andrea Vedaldi. Understanding deep image representations by inverting them. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 5188–5196, 2015.

[19] Sebastian Lapuschkin. Opening the machine learning black box with layer-wise relevance propagation. PhD thesis, Dissertation, Berlin, Technische Universität Berlin, 2018, 2019.

[20] Emre Vural, Mübeccel Çetin, Aytül Erçil, Gwen Littlewort, Marian Bartlett, and Javier Movellan. Automated drowsiness detection for improved driving safety. 2008.

[21] Dung Nguyen, Duc Thanh Nguyen, Sridha Sridharan, Simon Denman, Thanh Thi Nguyen, David Dean, and Clinton Fookes. Meta-transfer learning for emotion recognition. Neural Computing and Applications, pages 1–15, 2023.

[22] Hamidov Oybek Ikromovich and Babakulov Bekzod Mamatkulovich. Facial recognition using transfer learning in the deep cnn. Open Access Repository, 4(3):502–507, 2023.

[23] Prasannavenkatesan Theerthagiri. Stress emotion recognition with discrepancy reduction using transfer learning. Multimedia Tools and Applications, 82(4):5949–5963, 2023.

[24] Siti Khairuni Amalina Kamarol, Mohamed Hisham Jaward, Heikki Kälviäinen, Jussi Parkkinen, and Rajendran Parthiban. Joint facial expression recognition and intensity estimation based on weighted votes of image sequences. Pattern Recognition Letters, 92:25–32, 2017.

[25] Rabie Helaly, Seifeddine Messaoud, Soulef Bouaafia, Mohamed Ali Hajjaji, and Abdellatif Mtibaa. Dtl-i-resnet18: facial emotion recognition based on deep transfer learning and improved resnet18. Signal, Image and Video Processing, pages 1–14, 2023.

[26] Yupeng Sun, Hiroki Nomiya, and Teruhisa Hochin. Automatic evaluation of motion picture contents by estimation of facial expression intensity. In 2019 20th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD), pages 227–232. IEEE, 2019.

[27] Yamato Shinohara, Hiroki Nomiya, and Teruhisa Hochin. Estimation of facial expression intensity for lifelog videos retrieval. In 2018 5th International Conference on Computational Science/Intelligence and Applied Informatics (CSII), pages 133–138. IEEE, 2018.

[28] David Gunning, Mark Stefik, Jaesik Choi, Timothy Miller, Simone Stumpf, and Guang-Zhong Yang. Xai—explainable artificial intelligence. Science robotics, 4(37):eaay7120, 2019.

[29] Wojciech Samek, Alexander Binder, Grégoire Montavon, Sebastian Lapuschkin, and Klaus-Robert Müller. Evaluating the visualization of what a deep neural network has learned. IEEE transactions on neural networks and learning systems, 28(11):2660–2673, 2016.

[30] Amirata Ghorbani, Abubakar Abid, and James Zou. Interpretation of neural networks is fragile. In Proceedings of the AAAI conference on artificial intelligence, volume 33, pages 3681–3688, 2019.

[31] Andrey V Savchenko. Hsemotion team at the 6th abaw competition: Facial expressions, valence-arousal and emotion intensity prediction. arXiv preprint arXiv:2403.11590, 2024.

[32] Yini Fang, Liang Wu, Frederic Jumelle, and Bertram Shi. Integrating holistic and local information to estimate emotional reaction intensity. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 5934–5939, 2023.

[33] Patrick Lucey, Jeffrey F Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, and Iain Matthews. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In 2010 ieee computer society conference on computer vision and pattern recognition-workshops, pages 94–101. IEEE, 2010.

[34] Manuel G Calvo and Daniel Lundqvist. Facial expressions of emotion (kdef): Identification under different display-duration conditions. Behavior research methods, 40(1):109–115, 2008.

[35] MAH Akhand, Shuvendu Roy, Nazmul Siddique, Md Abdus Samad Kamal, and Tetsuya Shimamura. Facial emotion recognition using transfer learning in the deep cnn. Electronics, 10(9):1036, 2021.

[36] Ali Mollahosseini, Behzad Hasani, and Mohammad H Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. IEEE Transactions on Affective Computing, 10(1):18–31, 2017.

[37] Kai-Tai Song and Shuo-Cheng Chien. Facial expression recognition based on mixture of basic expressions and intensities. In 2012 IEEE International Conference on Systems, Man, and Cybernetics (SMC), pages 3123–3128. IEEE, 2012.

[38] Kuang-Yu Chang, Chu-Song Chen, and Yi-Ping Hung. Intensity rank estimation of facial expressions based on a single image. In 2013 IEEE International Conference on Systems, Man, and Cybernetics, pages 3157–3162. IEEE, 2013.

[39] Rohit Pal and C Satsangi. Facial expression recognition based on basic expressions and intensities using k-means clustering. Int. J. Sci. Res, 5:1949–1952, 2016.

[40] Jane Reilly Delannoy and John McDonald. Automatic estimation of the dynamics of facial expression using a three-level model of intensity. In 2008 8th IEEE International Conference on Automatic Face & Gesture Recognition, pages 1–6. IEEE, 2008.

SUDHEER BABU PUNURI is currently working in the Department of Computer Science, School of Engineering and Technology in the role of Assistant Professor. He has earned M.C.A (Master of Computer Applications) from D.L.R College, Affiliated to Andhra University and M.Tech from Swarnandhra College of Engineering affiliated to Jawaharlal Nehru Technological University (K). His research interest include Computer vision, Deep Learning and pattern recognition. Has co-authored and authored a few SCI journals, Scopus and international conferences.

DR. SANJAY KUMAR KUANAR is currently Heading the School of Computational Sciences in GIET University. He has earned his Ph.D. degree from Jadavpur University, Kolkata in November, 2015, obtained his M.E. (Computer Engineering) degree from Jadavpur University in 2007 and B. Tech Degree in Computer Science and Engineering from Utkal University in 1998. His research interests include computer vision, pattern recognition, multimedia computing and software metrics. He has authored and co-authored several publications in refereed SCI journals including IEEE Transactions on Multimedia, IEEE Transactions on Cybernetics, Elsevier's JVCIR and has many international IEEE peer-reviewed conference publications like ICPR, ICAPR etc. to his credit. His papers got many citations with h-index of 7 and i-10 index of 6. He is a reviewer for many national and international peer reviewed journals and member of technical program committee for many national and international conferences.

DR. TUSAR KANTI MISHRA is presently working as Associate Professor in the Dept. of Computer Science and Engineering at Manipal Institute of Technology Bengaluru. He has got more than 19 years of teaching experience. His research expertise is of ten years in the domain of Artificial Intelligence, Computer Vision, Pattern Recognition along with allied application areas. Dr. Tusar has availed the Erasmus Fellowship and worked under Prof. L R B Schomaker, the director of ALICE Lab, The Netherlands for a period of six months. He feels proud for being the students of famous research and academia personalities such as Dr. B. Majhi and Dr. Arun K Pujari. He has completed his Ph.D. from NIT Rourkela in 2015 along with MHRD fellowship. Dr. Tusar has been reviewer for multiple reputed journals. So, far he has filed four patents and one copyright to the IPR, Govt. of India and four more patents are in process. Dr. Tusar has more than 35 research publications and counting. Dr. Tusar has bagged research funding grant of INR 30 lakhs from SERB, Govt of India. He is also member of the BoS, Sambalpur University Inst. Of Inf. Tech. His qualification is B.Tech, M.Tech, Ph.D. He loves playing chess, table tennis, and cooking.

DR. SHIVA SHANKAR REDDY is working as Sr. Assistant Professor in Sagi RamaKrishnam Raju Engineering College, Bhimavaram, Andhrapradesh, INDIA. He has 16 Years Experience in Teaching and 10 Years of Experience in Research. He received his B.Tech. Degree in 2006 from JNTU in Department of Computer Science and Engineering. He received his M. Tech. Degree in 2009 from Andhra University in Computer Science and Technology and Completed his PhD Degree at Biju Patnaik University of Technology (BPUT), Rourkela, ODISHA, INDIA, on Febrauary 2024. His current research interests are Data Mining, Machine Learning, Medical Mining, Edge Computing, Cloud Computing, Image Processing & Computer Vision, Information Security and Algorithms. He has published 110+ Research Papers (Indexed in SCOPUS, SCIE, ESCI, etc...) in various International Journals and International Conferences. He is an editor for 4 Text Books and has written 10+ Book Chapters for various publishers. He has published 05 patents. He is acting as a Reviewer for various Peer Reviewed International Journals (SCI/SCOPUS/WOS) and International Conferences (IEEE/ Springer).

DR. VEERANKI V.R. MAHESWARA RAO is a leading Researcher and Academician in Department of Computer Science and Engineering and holds Ph.D. degree. He is currently working as a Professor in the Department of Computer Science and Engineering at Shri Vishnu Engineering College for Women (A), Andhra Pradesh, India. He is actively involved and successfully implemented three projects funded by DST. He has 45 research papers, 17 of which are Scopus-indexed and 7 of which are Web of Science-indexed. He has 23 years of experience that include 6 years of Industry experience, 19 years of Teaching experience and 15 years of Research experience. His Research interests include data mining, web mining, cloud computing, big data analytics, data science, artificial intelligence, and machine learning.

• • •