2023 International Conference on the Confluence of Advancements in Robotics, Vision and Interdisciplinary Technology Management (IC-RVITM) | 979-8-3503-0608-8/23/$31.00 ©2023 IEEE | DOI: 10.1109/IC-RVITM60032.2023.10435159

Facial Emotion Recognition in Unconstrained Environments through Rank-Based Ensemble of Deep Learning Models using 1-Cycle Policy

1st Sudheer Babu Punuri
Department. of CSE
GIET University, India
sudheerpunuri@giet.edu

2nd Sanjay Kumar Kuanar
Department. of CSE,
GIET University, India
sanjay.kuanar@giet.edu

3rd Tusar Kanti Mishra
Department of CSE
Manipal Institute of Technology Bengulur,
Manipal Academy of Higher Education,
Manipal, Karnataka, India.
tusar.mishra@manipal.edu

***Abstract*—The field of Facial Emotion Recognition (FER) has advanced considerably in the last few years. Much research relies on lab-controlled datasets, characterized by limitations in size, quantity, and quality. These datasets feature high-resolution static images captured in ideal conditions but lack fidelity in representing real-world scenarios. Hence, FER systems must be trained on primary data that includes real-world scenarios like facial expressions captured from various angles and in different lighting conditions, images with occlusion etc., broadly termed as unconstrained environment. To leverage the gap, this study emphasizes utilizing an AffectNet dataset that has samples close to real-world scenarios. In addition, we propose a novel ensemble framework to increase the accuracy of emotion recognition by harnessing the complementary strengths of three distinct deep-learning models: DenseNet169, EfficientNetB7 and InceptionV3. The key innovation lies in our novel ranking-based fusion technique, which introduces a unique perspective on model confidence and its relationship with prediction quality. The rank-based fusion approach optimally harnesses each base model's unique characteristics and strengths. Our experiments confirm the ensemble framework's effectiveness, outperforming individual models in facial emotion recognition.**

***Index Terms*—Facial Emotion Recognition, Ensemble Learning, Deep Learning, Ranks.**

I. INTRODUCTION

Within the dynamic field of human-computer interaction and affective computing, Facial Emotion Recognition (FER) has emerged as a pivotal and interdisciplinary field with profound implications for diverse applications. Connecting human emotional expressions with computational intelligence, Facial Emotion Recognition (FER) presents a hopeful path to improve the efficiency and effectiveness of diverse human-centric systems and technologies. At its core, FER entails the automated detection and analysis of human emotions through the interpretation of facial expressions. In addition to being essential for comprehending and addressing human emotions, this capacity has numerous uses in fields such as healthcare and entertainment to marketing and robotics. For instance, in healthcare [1] [2], FER can help in the early identification of patients' emotional discomfort, while in the entertainment industry, it can personalize user experiences by adapting content based on emotional states. Furthermore, in marketing [3], FER can be leveraged to gauge consumer reactions to products and advertisements, facilitating data-driven decision-making.

The inherent intricacy of Facial Emotion Recognition (FER) stems from the diverse aspects of human emotions and the variability in facial expressions, which are shaped by cultural, individual, and contextual factors. Consequently, the development of accurate and robust FER models poses a significant computational and scientific challenge. Throughout the years, researchers have delved into a multitude of approaches, ranging from traditional machine learning methods to state-of-the-art deep learning models, in their efforts to address this intricate challenge.

This research article delves into the realm of Facial Emotion Recognition, where we present an innovative approach to improving the accuracy of emotion recognition by employing an ensemble framework that combines the strengths of three deep learning models: DenseNet169, EfficientNet, and InceptionV3. Unlike traditional ensemble methods, our approach leverages a unique ranking-based fusion technique, which reimagines the relationship between model confidence scores and prediction quality. Through empirical evaluations, we demonstrate the effectiveness of our ensemble strategy in surpassing individual model performances, thus offering a promising solution for real-world emotion recognition applications, particularly in unconstrained environments. The comprehensive model is visually represented in Fig.1.

In the subsequent sections, we delve into the methodology, experimental results, and implications of our research, culminating in a comprehensive exploration of the benefits of rank-based ensemble methods in enhancing Facial Emotion Recognition systems. The contributions of the present research work encompass:

1) To focus on utilizing the AffectNet dataset, which contains samples that closely resemble real-world scenarios, improving the relevance and applicability of emotion recognition models.

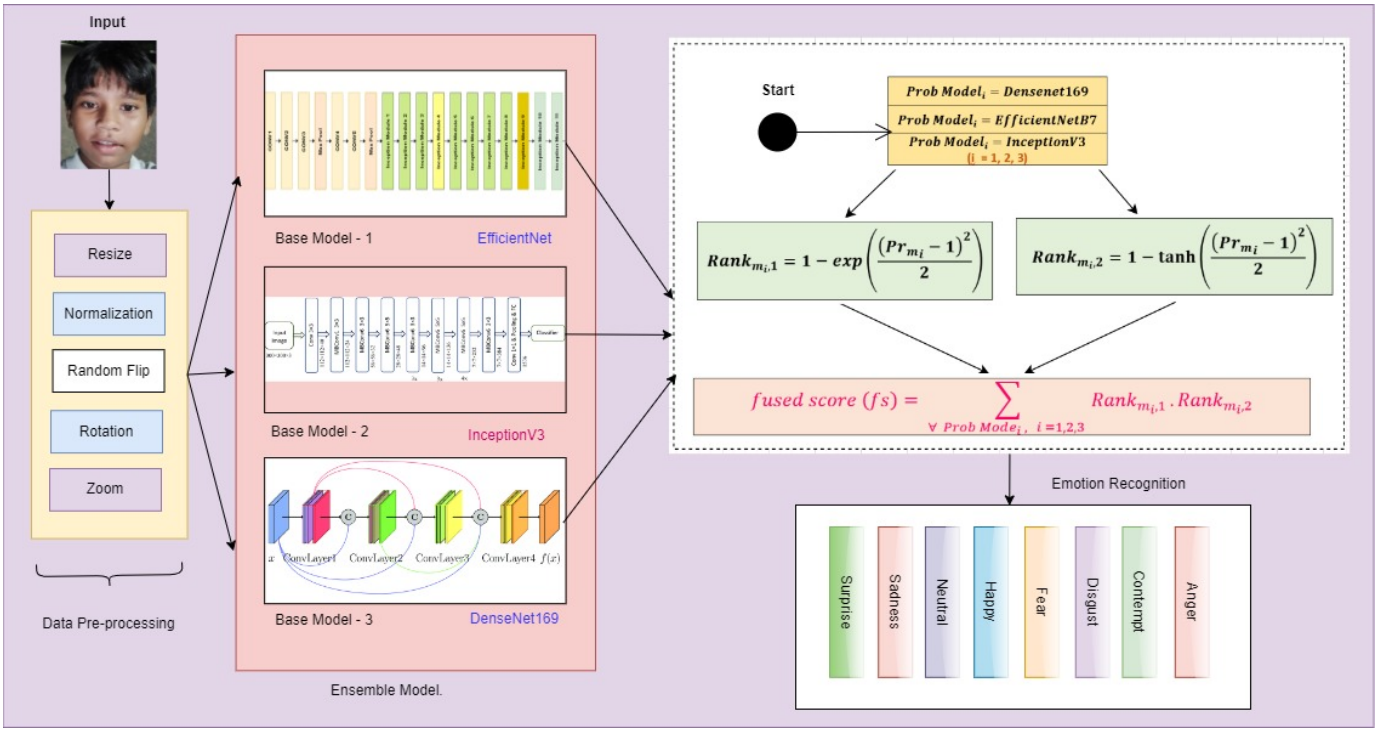

Fig. 1. Ensemble Model

2) To introduce a novel ensemble framework that leverages the strengths of three distinct deep-learning models, namely DenseNet169, EfficientNetB7, and InceptionV3, to enhance the accuracy of emotion recognition.
3) The proposed ensemble approach utilizes confidence scores to compute fuzzy ranks for classes employing two non-linear functions. The fused score is determined as the product of ranks produced by the three base classifiers. The predicted class corresponds to the one with a lower fused score.

II. LITERATURE SURVEY

In the literature, several methods and techniques are employed to resolve face emotion recognition challenges. The authors [18] proposed a model using a lightweight version of the convolutional neural network (CNN) and the AffectNet dataset. Their contribution also includes the steps required in uploading and storing an image for FER. A simple and efficient model, titled label distribution learning(LBL) [17] proposed. The ResNet was the architecture of this model. In another paper, [19] presents an adaptive multilayer perceptual attention network (AMP-Net) that takes its cues from the human visual system's facial features and its process for seeing faces. To unveil the inherent diversity and crucial information concerning facial emotions, AMP-Net extracts global, local, and salient emotional facial aspects by leveraging various fine-grained features. [4] has employed a multimodal fusion method that combines Temporal Convolutional Networks (TCN) and Transformer to enhance the performance of continuous emotion recognition. From the high-level and low-level features, these modules assist in removing the most pertinent and discriminative traits. To compute the probability score, the feature maps are then integrated and sent to the dense layers, which are followed by a softmax layer. [5] proposed a model incorporating various recurrent neural networks (LSTM, GRU, BiLSTM, and BiGRU) in conjunction with the foundational architecture of the EfficientNetV2-S convolutional neural network. The ensemble scheme fuses classifiers through a fuzzy rank-based [6] approach, using non-linear functions for decision-making. Authors [7] introduce EfficientNet-XGBoost, a novel method based on Transfer Learning. It combines EfficientNet and XGBoost, incorporating experimental enhancements to address challenges like vanishing gradients.

III. PROPOSED METHODOLOGY

Our proposed methodological framework is structured into several key components, each contributing to the overall research process. These components collectively form a comprehensive and systematic approach to our research work.

A. Overview of proposed ensemble method

In this section, a concise summary of the base learners and the customizations applied to them is provided. This is followed by the implementation of the fuzzy ranking process, which forms the core of this research. In this context, the aim of ensembling is to maximize the utilization of each confidence component generated by the base learners by mapping them into non-linear functions. One of the mapped values signifies the presence or proximity to 1, while the other signifies the

absence of 1. This suggested method solves the flaw in the traditional ranking techniques, which may produce inaccurate results if they fail to take into account the aforementioned fact [8] [9]. Our approach is tested on the human facial emotional image dataset (Affectnet) using three base learners in this work. First, we collect the confidence ratings after training the basic learners (a model customized with pre-trained models experimented on ImageNet [10]). Following this, we apply two separate functions with distinct concavities to map the scores, generating non-linear fuzzy ranks. The fusion of these two ranks yields a fused score, contributing to quantifying the overall deviation from the predicted outcome. A lower deviation indicates greater confidence in a certain class. The class with the minimum deviation value is assigned as the final class value and considered the winner. First, we briefly describe the CNN models that were pre-trained and utilized as foundation learners.

B. EfficientNet

EfficientNet [11] is a family of computationally efficient architectures for convolutional neural networks (CNNs) designed to balance model size and performance. It uses a compound scaling strategy to simultaneously scale up in terms of depth, width, and resolution to obtain innovative results on computer vision tasks. EfficientNet models use efficient building blocks, are pre-trained on large datasets like ImageNet, and strike a strong efficiency-accuracy tradeoff, making them popular for various computer vision applications. EfficientNet-B7 stands out as a highly efficient and effective deep learning model, owing to its state-of-the-art performance. With its impressive ability to generalize, low computational demands, and extensive community support, EfficientNet-B7 continues to be a top choice for image classification, object detection, and other computer vision applications, making it a valuable asset in the field of artificial intelligence.

C. DenseNet

DenseNet [13] represents a pioneering neural network architecture designed to overcome challenges in training deep convolutional neural networks (CNNs). Its hallmark feature is dense connectivity, where each layer is connected to every other layer in a feed-forward fashion. Dense blocks, consisting of multiple convolutional layers with batch normalization and ReLU activation, foster feature reuse and information flow. Transition layers manage feature map dimensions, while bottleneck layers optimize computational efficiency. DenseNet enables efficient training, mitigating the vanishing gradient problem, and has demonstrated versatility and state-of-the-art performance across various computer vision tasks, making it a prominent choice in deep learning research.

D. InceptionV3

InceptionV3 [12] is a deep convolutional neural network architecture renowned for its innovative inception modules. These modules employ multiple filter sizes within the same layer, enabling the network to capture features at various scales. InceptionV3 is characterized by its depth and intricate architecture, which promotes feature diversity and abstraction. It has achieved state-of-the-art results in image classification and other computer vision tasks. InceptionV3's success can be attributed to its ability to efficiently capture complex patterns and representations across different scales, making it a valuable asset in deep learning research and applications.

TABLE I
HYPERPARAMETERS USED DURING TRAINING THE BASE CLASSIFIERS.

Hyperparameters	Value(s)
Optimizer	AdamW
Weight Decay	0.001
Loss Function	Categorical Cross entropy
Batch Size	32 & 64
Dropout Rate	25
Regularization	0.001
Input Shape (Image)	$(72, 72, 3)$

E. Series of pre-trained models combined with custom layers.

To optimize the weights generated by pre-trained deep learning models, we enhance their structure by incorporating a few specialized layers. In Densenet169, EfficienNetB7, and Inceptionv3, each fully connected layer is expanded with an additional 1024, 1028, or 256 nodes beyond the existing architecture of these pre-trained models. To mitigate the vanishing gradient problem and accelerate learning, the Rectified Linear Unit (ReLU) activation function is utilized on these fully connected layers. Subsequently, to mitigate the overfitting 20% dropout layer is added. To avoid potential information loss from computing confidence scores directly from a substantial number of hidden units, this approach is implemented. The hyperparameters employed for training these models have been determined through thorough experimentation and are detailed in Table I The weights of the other layers are kept constant throughout training, and training is limited to the modified layers that were introduced to these CNN models. Since the model weights were pre-trained on the ImageNet data, they are already optimized for image classification, hence the number of epochs is set to 50.

F. 1 - Cycle Policy

The 1-cycle policy [15] involves the learning rate transitioning from an initial value to a high rate, followed by a decrease to a minimum rate significantly lower than the initial learning rate. This concept was initially introduced in Super-Convergence [16]. A cycle can be explained using either of these two methods: a reward for reaching a specified number of steps is explicitly defined, considering a few epochs and a set number of steps for each epoch. The instances of incremental steps are evaluated in this scenario, as depicted in the following equation.

$$T_{steps} = epochs * S_{epochs} \tag{1}$$

where T_{steps} is Total steps.
S_{epochs} is Steps per epoch. The idea behind the 1-cycle policy is to start with a relatively low learning rate, gradually increase

it to a high value, and then anneal it back down to lower value. This approach helps in several ways:

1) **Faster Convergence**: It allows the model's convergence faster during the initial phase of training when the learning rate is low.
2) **Escape Saddle Points**: The high learning rate phase can help the model escape saddle points or poor local minima.
3) **Regularization**: The cyclical nature of learning rates acts as a form of regularization, preventing the model from overfitting.
4) **Improved Generalization**: By cycling the learning rate, the model can potentially generalize better to unseen data.

G. Mathematical Foundation of the Ensemble Model Construction

We present the use of a mathematical approach for ensemble modeling in this section. To implement into action an ensemble methodology that uses three deep learning architectures(pre-trained) as base learners. These models provide confidence scores for the AffectNet dataset, specifically for a set of C distinct classes. The confidence score is represented by each base learner i as $Pr_1^i, Pr_2^i, Pr_3^i, ..., Pr_C^i$ where $i \in \{1, 2, 3, ..\}$. The variable $'i'$ serves as an analyzer of these 3 base models which is represented by Eqn. (2).

$$\sum_{k=1}^{C} Pr_k^i = 1, \forall\, i \in \{1, 2, 3\} \tag{2}$$

Let

$$(Rank_1^{i_1}, Rank_2^{i_1}, Rank_3^{i_1}, \ldots, Rank_C^{i_1})$$

and

$$(Rank_1^{i_2}, Rank_2^{i_2}, Rank_3^{i_2}, \ldots, Rank_C^{i_2})$$

are the fuzzy ranks produced by the 2 functions which are non-linear. These non-linear functions help in generating fuzzy ranks in this proposed ensemble model. They are exponential and hyperbolic tangent functions denoted by Eqn.(3) and Eqn.(4). Here, the variable x represents the class's probability for a sample image in our ensemble model.

$$y = 1 - exp\left(-\frac{(x-1)^2}{2}\right) \tag{3}$$

$$y = 1 - tanh\left(\frac{(x-1)^2}{2}\right) \tag{4}$$

Inorder to calculate fuzzy ranks in our ensemble model the x is replaced with Pr_k^i. Eqn. (5) and Eqn. (6). are the new equations for generating Ranks.

$$Rank_k^{i_1} = 1 - tanh(\frac{(Pr_k^i - 1)^2}{2} \tag{5}$$

$$Rank_k^{i_2} = 1 - exp(-\frac{(Pr_k^i - 1)^2}{2}) \tag{6}$$

1) For this study, $exp(-\frac{(x-1)^2}{2})$ has a downward concavity in the interval [0,1] in our study. It's crucial to highlight that the negative gradient of the function in the interval [0, 1] causes the rank score output to incline toward 1, indicating an upward concave shape for the function.
2) For this study, $tanh(\frac{(x-1)^2}{2})$ is concave towards upward in its domain of definition [0, 1] in our study. This function has a +ve gradient in a specific range, which means that the function is increasing in that interval. This aligns with the idea that the output rank score is increasing and moving closer to 0.

.

The functions that compute non-linear rankings will have a domain of definition of [0, 1] as $P_k^i \in [0, 1]$. Eqn.(5) offers a classification reward. The value of Equation (2) rises as x gets closer to 1, meaning that the reward is increased. In contrast, if x approaches ≤ 0, the divergence will be greater for Eqn. (6) when calculating the departure from 1. Let

$$(RS_1^i, RS_2^i, RS_3^i, \ldots, RS_C^i)$$

be the fused rank scores, where RS_k^i is derived from the following Eqn.(7)

$$RS_k^i = R_1^{i_1} X R_1^{i_2} \tag{7}$$

The rank score is calculated by multiplying the reward by the variance for a specific confidence score that was acquired from a baseline learner. Since the range of Eqn. (3) is smaller than the range of Eqn. (2), Eqn.(3) will control the characteristics of the final product. A lower rank score is implied by a smaller divergence determined from the confidence score. Ultimately, the sole factor taken into account for determining the fused scores are the rank scores. Since the RS_k^i is the result of fuzzy ranks produced by the two different types of functions, it will represent the degree of confidence towards a specific class. Since FS_k is determined by Eqn.(8), the fused score tuple is now $FS_1, FS_2, FS_3, \ldots, FS_C$.

$$FS_k = \sum_{i=1}^{L} RS_k^i, \forall\ k = 1, 2, 3, \ldots, C \tag{8}$$

The final score for every class can be obtained by fusing the scores together. The class with the lowest fused score is then determined by applying Eqn.(9) to determine the winner. The fusion strategy's computational complexity is O(number of classes).

$$Emotion\_class(I) = \min_{\forall k} FS_k \tag{9}$$

The mathematics involved in the ensemble implementation of the model is shown in Fig. 2. For both Eqn.(3) and Eqn.(4), a nature of the demonstrates unambiguously that an increase in probability leads to a lower ultimate rank. In this way, correctness can be demonstrated.

H. Dataset

AffectNet [14] is a vast and diverse dataset used for facial emotion analysis, featuring over 1 million facial images captured from various sources. It covers eight emotion categories, including neutral, happiness, sadness, surprise,

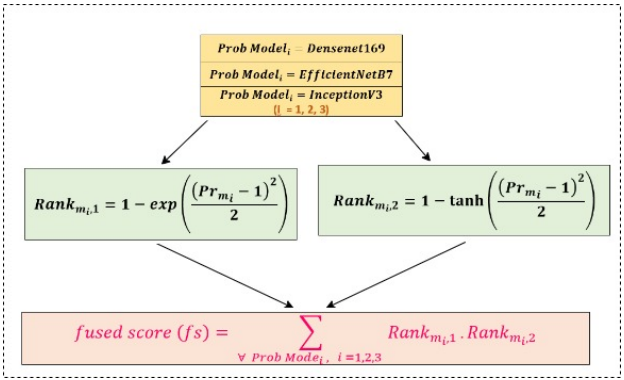

Fig. 2. Fused Score generation for the 3 base classifiers

fear, disgust, anger, and contempt. These annotations provide valuable training and evaluation data for emotion recognition models. AffectNet poses challenges due to its "in the wild" nature, with data collected from the internet and social media, requiring careful preprocessing. Researchers employ this dataset to develop deep learning models for facial emotion recognition, benefiting applications like sentiment analysis and affective computing. AffectNet is publicly available, serving as a valuable resource in the field of computer vision and emotion analysis research. The distribution of images in emotional classes is shown in Fig 4. The annotations also introduce challenges due to potentially mislabeled or inaccurately annotated images. The dataset's sample images are illustrated in Fig. 5

I. Evaluation Metrics

We have employed the four widely-used evaluation metrics of Accuracy, Precision values, Recall values, and F1-Score to verify the performance of the suggested model.

$$Accuracy = \frac{Numberofcorrectprediction}{Totalnumberofpredictions} \tag{10}$$

Precision: Precision is used to evaluate the performance of a classification model. It measures the accuracy of the model's positive predictions, indicating the proportion of correctly identified positive cases among all instances predicted as positive. It is particularly important in situations where false positives are costly or where we want to ensure that the positive predictions are highly reliable.

$$Precision = \frac{TP}{TP + FP} \tag{11}$$

Recall: Recall, in the view of model evaluation, measures a model's ability to detect and include all actual positive cases within a dataset.It represents the proportion of positive instances accurately identified by the model in relation to the overall number of positive instances present in the dataset. A high recall value suggests that the model excels in minimizing false negatives, ensuring it captures as many positive cases as possible, even if it results in a few false positives. The recall is particularly vital in situations where missing a positive case could have significant repercussions, such as in medical diagnoses or fault detection systems.

$$Recall = \frac{TP}{TP + FN} \tag{12}$$

F1-Score: The F1-Score is the harmonic mean of precision and recall. It considers both false positives and false negatives, making it a robust metric for evaluating a model's capability to accurately classify instances of a particular class while minimizing incorrect predictions. A high F1-Score indicates a model that achieves both high precision and high recall, effectively balancing the trade-off between false positives and false negatives.

IV. Results and Discussion

A. Experimentation

This evaluation process allowed us to pinpoint scenarios or data points where the ensemble model exhibited exceptional predictive power. By scrutinizing these instances, we gained valuable insights into the strengths and weaknesses of the model, enabling us to make informed decisions regarding its deployment or potential improvements. Fig.6 provides a visual representation of our experimental outcomes. Fig. 3 shows the ensemble process with the confidence score, ranks, fused ranks and fused score for the final prediction class. The confidence score of the 3 base learners for an image that belongs to the Disgust emotion class is taken, which is shown in Fig.3(a) - (c). The confidence score (probability) given by the Densenet model of the Anger class is 0.8387 so ranks are 0.9871 and 0.0130 which produces a fused rank of 0.0128. Similarly, all the fused ranks for all emotions are calculated. Out of the fused rank for all 8 emotions, 0.0128 is the minimum value hence, an emotion corresponding to 0.0128 is treated as the predicted emotion which is anger. Similarly fused ranks for efficientnet and inception are also calculated. The fused ranks from all the three base learners are considered which are shown in Fig.3(d). Clearly from Fig.3(d) Densenet, Efficientnet and inception predict anger, disgust and contempt respectively, whereas Densenet and inception are wrongly predicted. Now using the fused ranks from the base learners a fused score (last column of 3(d)) is evaluated. The fused score has a minimum of 0.0002. The matching emotion is disgust, which is regarded as the ultimate winner class which is the actual label of the image. This becomes the robust decision of the model.

Experimentation is done with Google Colab Pro using the configuration that includes a Tesla T4 GPU with 15,360 MiB of GPU memory, Python version 3.10.3, and Tensorflow version 2.13.0.

B. Confusion Matrix

Confusion matrix is a crucial tool for evaluating the performance of emotion recognition models on the test dataset. This matrix provides a detailed breakdown of predicted versus actual emotion classes, shedding light on the model's strengths and weaknesses. In the context of AffectNet's eight emotion classes, confusion matrix allows us to visualize not only

Fig. 3. (a) through (c) shows the Fused Ranks for the base classifiers namely, DenseNet, EfficientNet and Inception. (d) is the Fused Score corresponding to the winner class.

TABLE II
PERFORMANCE METRICS FOR INDIVIDUAL BASE CLASSIFIER AND AN ENSEMBLE MODEL.

Emotion	Densenet			EfficientNet			Inception			Ensemble		
	Precision	Recall	F1_score	Precision	Recall	F1_score	Precision	Recall	F1_score	Precision	Recall	F1_score
Anger	0.84	0.88	0.86	0.84	0.86	0.85	0.47	0.84	0.6	0.83	0.84	0.95
Contempt	0.9	0.87	0.88	0.91	0.91	0.91	0.7	0.89	0.78	92	0.89	0.91
Disgust	0.89	0.84	0.86	0.85	0.84	0.84	0.76	0.81	0.76	0.87	0.84	0.87
Fear	0.9	0.82	0.86	0.83	0.87	0.85	0.77	0.61	0.68	0.89	0.85	0.88
Happy	0.98	0.97	0.97	0.98	0.97	0.98	0.97	0.96	0.96	0.97	0.96	0.96
Neutral	0.95	0.98	0.96	0.96	0.97	0.96	0.95	0.92	0.93	0.95	0.95	0.93
Sadness	0.84	0.87	0.85	0.86	0.85	0.85	0.64	0.62	0.63	0.78	0.81	0.80
Surprise	0.86	0.88	0.87	0.89	0.84	0.87	0.77	0.77	0.77	0.79	0.89	0.91
Accuracy	90%			91.01%			75.60%			**92.32%**		

correct classifications but also instances of misclassification. The confusion matrix for the base learners is displayed in Fig. 7, Fig. 8 and Fig. 9. Confusion matrix is an 8 x 8 matrix with 8 emotion classes. From Fig.7, 8 the performance of the model is evident from the diagonal values, which represent accurately predicted image that belongs to 8 emotions. Fig. 9 is slightly having little less performance, however in the ensemble model promising results have been produced.

TABLE III
ACCURACY COMPARISON WITH THE EXISTING WORKS..

Methodology used	Accuracy (%)
Label distribution learning (LDL) [17]	63.70
A Light weighted CNN [18]	77
Adaptive multilayer perceptual attention network (AMP-Net) [19]	64.54
Handcrafted and local learning [20]	59.58
Ensemble of pre-trained models (Proposed Model)	91.34

C. Comparison to State-of-the-art

To validate the performance of the ensemble model, we conducted experiments using a subset of samples from the test dataset where the model demonstrated superior performance. In these experiments, we aimed to identify specific instances

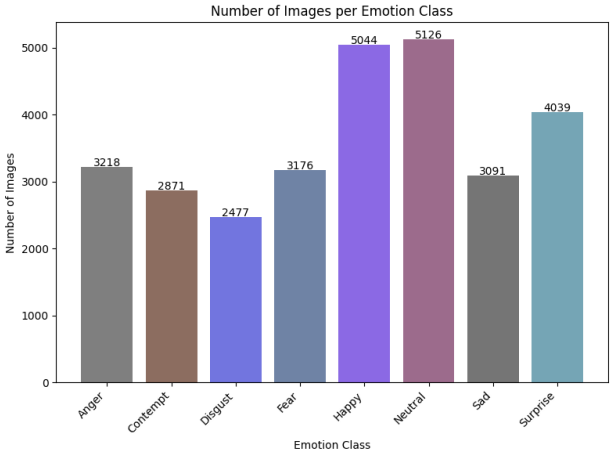

Fig. 4. Sample distribution per emotions in AffectNet Dataset.

Fig. 5. AffectNet dataset samples for 8 emotion classes.

where the ensemble model excelled. 4357 images from the Affectnet dataset's eight emotion classes are included in the test split of the dataset. Currently, all eight emotion classes' random images are being used to test the ensemble model. The results show outstanding potential. Fig 10 displays a few examples of images that were evaluated and accurately predicted. Compared to the state of the art, our suggested

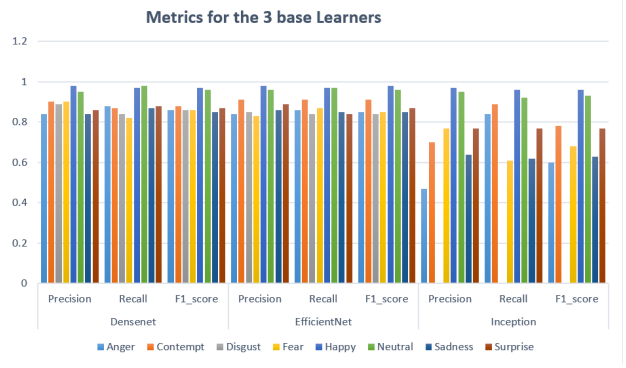

Fig. 6. Evaluation metric for 3 base learners.

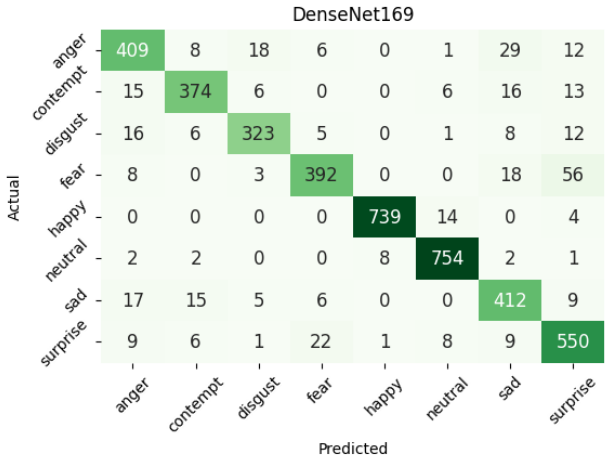

Fig. 7. Confusion matrix for the base model DenseNet169

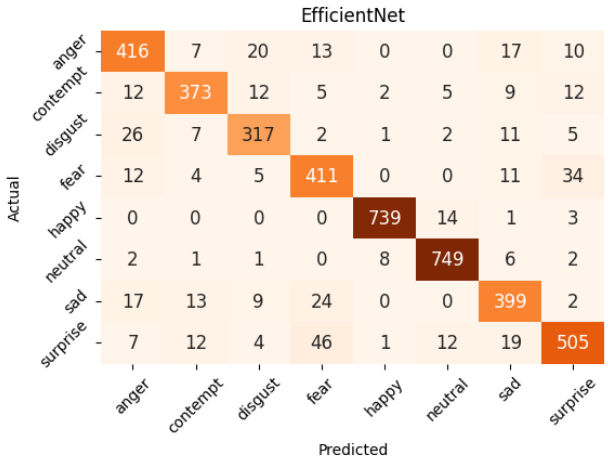

Fig. 8. Confusion matrix for the base model EfficientNetB7

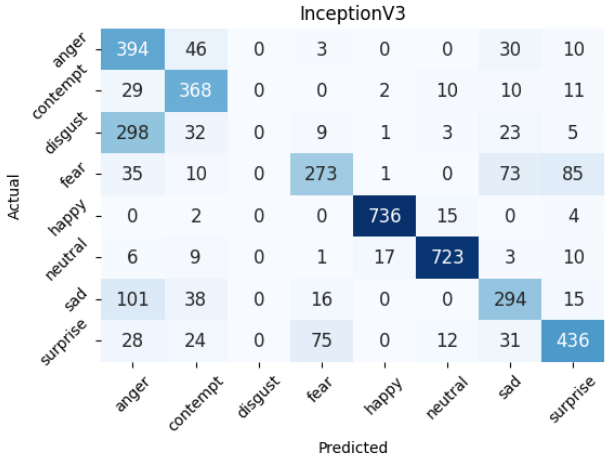

Fig. 9. Confusion matrix for the base model InceptionV3

ensemble model performs better however, some images are incorrectly identified by the ensemble model. Fig.11 displays

Fig. 10. Sample images from all 8 emotion classes evaluated and predicted correctly.

Fig. 11. Wrongly predicted sample images by the ensemble model.

the example image. The primary issue with the Affectnet dataset is a situation of class imbalance, which can be taken care of by applying the GAN technique for higher efficiency. Authors [17] introduced, a label distribution learning (LDL) method for training, achieving state-of-the-art accuracy on datasets with occlusions, pose variations, and benchmark datasets such as RAF-DB, CAER-S and AffectNet. Affectnet accuracy is 63.70%. [18] has experimented on Afffecnet using a robust CNN model with an accuracy of 55.09%. AMP-Net [19] framework has been proposed and experimented on 3 popular benchmark datasets. The model outperforms with an accuracy of 64.54%. Occlusion and changing stances result in inadequate facial information, notwithstanding AMP-Net's resilience to various facial situations. The Table. III illustrates a comparison between our proposed model and existing works.

V. Conclusion and Future Scope

In conclusion, our study illuminates the dynamic landscape of Facial Emotion Recognition (FER) and underscores the pressing need for adaptable systems that can excel in real-world, unconstrained scenarios. Focused on leveraging the AffectNet dataset, mirroring authentic conditions, our research aims to elevate the precision and practicality of FER systems. The introduction of a novel ensemble framework, amalgamating the strengths of Deep Learning models—DenseNet169, EfficientNetB7, and InceptionV3—has resulted in a substantial improvement in emotion recognition accuracy, reaching an impressive 91.34%. This significant enhancement underscores the efficacy of our approach, emphasizing its potential to advance the field of FER. The novel ranking-based fusion technique introduced in our study further bolsters model confidence, influencing prediction quality, and unifying decision-making processes. Looking forward, our ongoing efforts involve extending this methodology to the FER2013 dataset, ensuring broader applicability and reliability across diverse data sources and real-world contexts.

References

[1] Mosquera, Candelaria, et al. "Introducing Computer Vision into Healthcare Workflows." Digital Health: From Assumptions to Implementations. Cham: Springer International Publishing, 2023. 43-62.
[2] Paulauskaite-Taraseviciene, Agne, et al. "Geriatric Care Management System Powered by the IoT and Computer Vision Techniques." Healthcare. Vol. 11. No. 8. MDPI, 2023.
[3] Zhang, Shunyuan, and Kannan Srinivasan. Marketing Through the Machine's Eyes: Image Analytics and Interpretability. Vol. 20. Emerald Publishing Limited, 2023.
[4] Zhou, Weiwei, et al. "Continuous emotion recognition based on tcn and transformer." arXiv preprint arXiv:2303.08356 (2023).
[5] Oucherif, Sabrine Djedjiga, et al. "Facial Expression Recognition Using Light Field Cameras: A Comparative Study of Deep Learning Architectures." 2023 IEEE International Conference on Image Processing (ICIP). IEEE, 2023.
[6] Manna, Ankur, et al. "A fuzzy rank-based ensemble of CNN models for classification of cervical cytology." Scientific Reports 11.1 (2021): 14538.
[7] Punuri, Sudheer Babu, et al. "Efficient net-XGBoost: an implementation for facial emotion recognition using transfer learning." Mathematics 11.3 (2023): 776.
[8] Monwar, Md Maruf, and Marina L. Gavrilova. "Multimodal biometric system using rank-level fusion approach." IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) 39.4 (2009): 867-878.
[9] Abaza, Ayman, and Arun Ross. "Quality based rank-level fusion in multibiometric systems." 2009 IEEE 3rd International Conference on Biometrics: Theory, Applications, and Systems. IEEE, 2009.
[10] Deng, Jia, et al. "Imagenet: A large-scale hierarchical image database." 2009 IEEE conference on computer vision and pattern recognition. Ieee, 2009.
[11] Koonce, Brett, and Brett Koonce. "EfficientNet." Convolutional Neural Networks with Swift for Tensorflow: Image Recognition and Dataset Categorization (2021): 109-123.
[12] Szegedy, Christian, et al. "Rethinking the inception architecture for computer vision." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.
[13] Zhang, Ke, et al. "Multiple feature reweight densenet for image classification." IEEE Access 7 (2019): 9872-9880.
[14] Mollahosseini, Ali, 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 (2017): 18-31.
[15] Lo, S-CB, et al. "Artificial convolution neural network techniques and applications for lung nodule detection." IEEE transactions on medical imaging 14.4 (1995): 711-718.
[16] Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." Advances in neural information processing systems 25 (2012).
[17] Zhao, Zengqun, Qingshan Liu, and Feng Zhou. "Robust lightweight facial expression recognition network with label distribution training." Proceedings of the AAAI conference on artificial intelligence. Vol. 35. No. 4. 2021.
[18] Siddiqui, Nyle, et al. "A robust framework for deep learning approaches to facial emotion recognition and evaluation." 2022 Asia Conference on Algorithms, Computing and Machine Learning (CACML). IEEE, 2022.
[19] Liu, Hanwei, et al. "Adaptive multilayer perceptual attention network for facial expression recognition." IEEE Transactions on Circuits and Systems for Video Technology 32.9 (2022): 6253-6266.
[20] Georgescu, Mariana-Iuliana, Radu Tudor Ionescu, and Marius Popescu. "Local learning with deep and handcrafted features for facial expression recognition." IEEE Access 7 (2019): 64827-64836.