EAI Endorsed Transactions
on AI and Robotics
Research Article EAI.EU

Facial Emotion Recognition by CNN Combined Ensemble Model

Kurbanov Abdurahmon Alishboyevich[1*]

[1]PhD student, Jizzakh branch of the National University of Uzbekistan named after Mirzo Ulugbek, Jizzakh, 130100, Uzbekistan.

Abstract

Studying emotions can provide important information about a person's mental state. According to research, more than 50% of a person's current emotions can be identified from the human face. In this research, we propose an ensemble model for emotion recognition from facial images, which is obtained by combining the results obtained by retraining previously trained convolutional neural networks on a new and high-quality FaceEmocDS dataset. The methodological advantage of the ensemble model we propose is that the combination of VGG19, ResNet50, and DenseNet121 models allows us to take advantage of the strengths of each architecture: the ability to extract detailed features of VGG19, the stable learning process of ResNet50 through residual connections, and the efficiency of feature reuse of DenseNet121. This approach improves the results of individual models, increasing the accuracy to 85.66% . The FaceEmocDS dataset consists of 72,412 images and includes eight emotion classes, including a unique "contempt" class. The results show significant superiority when compared to other datasets (FER2013, AffectNet, CK+) and studies.

Keywords: CNN, ResNet, VGGNet, DenseNet, transfer learning, ensemble model, ESPCN, hash_compare.

Received on 14 July 2025, accepted on 06 September 2025, published on 12 September 2025

doi: 10.4108/airo.9720

*Corresponding author. Email: mr.kurbanov144@gmail.com

1. Introduction

Facial emotion recognition is an important tool in the field of affective computing to analyze human mental states and improve human-computer interaction. In recent years, deep learning, especially convolutional neural networks, has made significant progress in this field. However, existing datasets (e.g., FER2013, CK+, AffectNet) have limitations such as low-quality images, class imbalance, and lack of adaptability to real-world conditions. To overcome these problems, the FaceEmocDS dataset was developed, which contains 72,412 high-quality images (224x224 RGB, standardized with ESPCN neural network) and 8 emotion classes (anger, contempt, disgust, fear, happy, neutral, sad, surprise).

In our study, we tested the pre-trained VGG19, ResNet50, DenseNet121 and ViT-B/16 models on the FaceEmocDS dataset, achieving accuracy results of 75.92% for ResNet-50, 74.50% for VGGNet-19, 76.10% for DenseNet-121 and 74.71% for ViT-b16, respectively. These results were obtained by training the above pre-trained CNN models on the FaceEmocDS dataset for 30 epochs, and the best performing models were saved as a file. While these results confirm the quality and diversity of the dataset, individual models faced limitations in detecting subtle emotions (e.g., "contempt"). Therefore, in this paper, we proposed an ensemble model that combines the strengths of the VGG19, ResNet50 and DenseNet121 CNN models. The proposed ensemble model was stored in *.pth files representing the best training state for each model. In this study, these best results and neural network weights were combined and retrained on the FaceEmocDS dataset to build an ensemble model. The training was conducted for 15 epochs and achieved an accuracy of 85.66%. The results were compared with other studies and the relevance of the model in real-world applications was discussed.

2. Related works

Facial Expression Recognition and its more nuanced subfield, micro-expression recognition, represent significant challenges in computer vision and affective computing. The complexity of facial affect, influenced by factors such as pose, occlusion, illumination, and individual subjectivity, necessitates robust and sophisticated computational approaches. Recently, ensemble learning techniques, which combine multiple models to improve generalization and accuracy, have been extensively applied to these tasks. This review synthesizes current research on ensemble-based methods for Facial Expression Recognition and Micro-Expression Recognition, highlighting the architectural innovations and performance outcomes reported in the literature.

A prominent approach involves creating multi-stream deep learning architectures to capture diverse facial features. Perveen et al. [1] proposed a multi-stream deep convolutional neural network for MER, integrating features from ResNet, DenseNet, and VGG architectures. To manage the high dimensionality of these features, Principal Component Analysis (PCA) was employed for reduction. A stacking ensemble classifier, utilizing Random Tree, J48, and Random Forest as base learners and a Random Forest meta-learner, was implemented. Evaluated on the CASME-II, CASME2, SMIC, and SAMM datasets, the proposed method was compared against twelve existing approaches. The results demonstrated superior performance in both accuracy and computational efficiency, establishing the effectiveness of multi-stream feature fusion with ensemble classification.

Transfer learning, combined with ensemble strategies, has proven highly effective. Almubarak and Alsulaiman [2] fine-tuned a pre-trained EfficientNet-B0 model on a dataset of grayscale images across eight emotion classes. Their methodology incorporated transfer learning and a stacking ensemble with binary classifiers and a meta-classifier. This approach achieved remarkable results, reporting 100% accuracy on their test set and 92% accuracy on the standardized Cohn-Kanade (CK+) dataset, underscoring the potential of combining strategic fine-tuning with ensemble learning for high-accuracy FER.

Innovative single-model architectures enhanced with ensemble-based feature fusion and decision strategies have also been explored. Zhou, Xie, & Tian introduced a ResNet18-based model (R18+FAML) that integrated multiple loss functions and attention blocks to improve feature diversity [3]. They further proposed a Genetic Algorithm-based feature fusion (FGA) and a Top-Two Voting (T2V) ensemble strategy, creating the R18+FAML-FGA-T2V model. The ensemble model yielded high accuracy scores across multiple benchmarks (reported as 91.59, 63.27, and 66.63 on different datasets), outperforming the base single model and demonstrating the value of evolutionary algorithms in feature synthesis for ensemble construction.

Beyond visual data, ensemble methods excel in multimodal emotion recognition by fusing heterogeneous sensor data. Younis et al. [4] collected a multimodal dataset comprising environmental, physiological, and emotional response data. They evaluated various ensemble methods (Bagging, Boosting, Stacking) using KNN, Decision Tree, Random Forest, and SVM as base learners, with a Decision Tree meta-learner for stacking. Their results indicated that the stacking technique achieved the highest accuracy of 98.2%, compared to 96.4% for Bagging and 96.6% for Boosting, validating stacking as the most effective method for integrating sensor data to create a subject-independent emotion recognition model.

The strategic construction of ensembles is critical for performance. Renda et al. [5] conducted a comprehensive study on ensemble strategies specifically for deep learning in FER. They evaluated the impact of different sources of variability, aggregation schemes, and the number of base classifiers. A key finding was that preprocessing and pre-training procedures provided sufficient base classifier diversity. They also concluded that increasing ensemble size yields diminishing returns beyond a certain threshold, advising against excessively large ensembles.

Pandit, D. and Jadhav, S. [6] presented a balanced prediction method for all major facial emotions despite age and occlusion. The real-time facial emotion prediction methodology using an ensemble classifier, incorporating deep CNN models as the primary base classifier and addressing the problem of unbalanced datasets, is presented. The CK+ and JAFFE datasets are synthetically improved through image augmentation approaches. A metaclassifier using a combination of majority and relative voting methods is applied at level 2 to improve the accuracy of individual emotions. The proposed method is tested using randomly selected facial expression images from the internet and improves the overall accuracy.

The theoretical underpinnings of these approaches are covered in reviews such as that by Ganaie et al. [7] who provided a comprehensive overview of deep ensemble models, categorizing them by techniques like boosting, stacking, negative correlation, and heterogeneous ensembles, and discussing their applications across various domains.

Comparative studies consistently affirm the superiority of ensemble methods. Sağbaş, Uğur, and Korukoğlu evaluated ensembles using Bayesian Networks, k-NN, and Random Forest base learners combined via Adaboost, Bagging, Random Subspaces, and Voting [8]. The best result, an accuracy of 91.11%, was achieved by the Random Subspaces ensemble with a Random Forest classifier, confirming that ensembles significantly boost classification success over single models.

Optimizing the ensemble itself can lead to further improvements. Choi and Lee formulated the combination of Deep Convolutional Neural Network (DCNN) predictions as a stochastic optimization problem, using simulated annealing to find optimal ensemble weights that minimize generalization error. Evaluated on challenging "in-the-wild" datasets (FER2013, SFEW2.0, RAF-DB), their ensemble achieved competitive accuracies of 76.69%, 58.68%, and 87.13%, respectively [9].

Heterogeneous ensemble techniques are particularly powerful for multimodal data. Esfar-E-Alam et al. [10]

employed six models for audio and text modalities, combined using hard voting, soft voting, blending, and stacking. Their results showed that stacking was the most effective technique, achieving a weighted accuracy of 81.2% for 4-class emotion recognition on the IEMOCAP dataset, outperforming existing methods.

For complex tasks like multi-label FER, novel frameworks are being developed. Li, Luo, Zhang and Huang proposed a Multi-Feature Joint Learning Ensemble (MF-JLE) framework that combines global features with local key features. Their framework incorporated ensemble learning into the architecture itself, using a joint loss function for iterative optimization. This design improved multi-label recognition accuracy by treating different feature modules as weak classifiers within the ensemble [11].

Temporal information in videos can be leveraged through ensemble methods. Nguyen et al. [12] proposed a two-step method where spatial features were first extracted from each frame and then treated as temporal data for sequence-based classification. By ensemble connections within a convolutional network, they achieved superior results on the FER2013 dataset compared to state-of-the-art methods at the time.

For robust video emotion recognition, Smitha, Sendhilkumar, and Mahalakshmi developed three parallel CNNs for different detection and tracking methods (HOG-KLT, Haar-SVM, Patch-based). An ensemble of these networks achieved a high detection accuracy of 92.07% on videos containing both occluded and non-occluded faces [13].

Applied studies demonstrate the practicality of these methods. Muhajir et al. [14] used an ensemble of ResNet, MobileNet, and Inception to classify emotions of school students in Indonesia. Their approach achieved a high precision, recall, and F1-score of 90%. Similarly, Gupta, Kumar, & Tekchandani (2023) created an ensemble of VGG19 and ResNet50 via transfer learning to monitor student cognitive states (attention/inattention) in online learning. Their system achieved high recognition rates (93.11%, 92.34%, 91.12%) on new datasets, surpassing existing method performance.

Gupta et al. [15] proposed the EDFA framework to monitor students' cognitive states (attention, inattention) in adaptive online learning environments using ensemble deep CNNs. Three models (FT-EDFA, FC-EDFA, OT-EDFA) were developed with transfer learning on VGG19 and ResNet50, achieving recognition rates of 93.11%, 92.34%, and 91.12% on a custom dataset, outperforming existing methods. The system provides real-time feedback to instructors for adaptive teaching.

In a research paper published by Dương, Hải et al., multi-level features in a convolutional neural network were used to detect facial expressions [16]. Based on the researchers' observations, they introduced various network connections to improve the classification task. Combining the proposed network connections, they achieved competitive results compared to state-of-the-art methods on the FER2013 dataset.

The ensemble model proposed by Erwin Moung et al. [17] for emotion recognition from facial images combines three main convolutional neural networks, namely Custom CNN, ResNet50, and InceptionV3. In the study, the average ensemble classifier method of the model was used to combine the predictions from the three models. Then, the proposed facial expression recognition model was trained and tested on a dataset with an uncontrolled environment. The experiment showed that the aggregation of multiple classifiers was superior to all single classifiers in classifying positive and neutral expressions. However, only the ResNet50 model was said to be the best choice in classifying disgust, anger, and sadness.

The ensemble model for emotion recognition from facial images proposed by Viola Bakiasi and Markela Muça consists of three advanced convolutional neural networks: Inception V3, ResNet50, and SPP-net (Spatial Pyramid Convolution Networks), which are trained on the AffectNet dataset [18]. The ensemble model achieves 85.2% classification accuracy, which is 5-10% better than the best individual CNN. This shows the potential of ensemble methods in the field of facial expression recognition, showing a significant improvement. The paper highlights future research directions, including studying end-to-end trained ensemble models and collecting different datasets to further refine and improve the model performance.

The ensemble classifier model proposed by J.X. Chang et al. [19] for emotion recognition from facial images includes four models: VGG-19, VGGFace, ViT-B/16, and ViT-B/32. The model was tested on three datasets with clean data. The accuracy of the results was 100% on the clean CK+ dataset, 76.30% on the clean FER-2013 dataset, and 100% on the clean JAFFE dataset, respectively.

A. Das et al. [20] proposed combining previously trained versions of DenseNet CNN models into an ensemble model. This study was conducted to determine how DenseNet networks (DenseNet121, DenseNet169, DenseNet201) perform in emotion recognition from the FER2013 dataset. This approach involves training these models separately and then building an ensemble model. In this study, DenseNet121, DenseNet169, DenseNet201 and their ensemble model achieved accuracies of 71.59%, 72.01%, 72.32% and 74.16% , respectively . The study results showed that the ensemble model (DenseNet121 + DenseNet169 + DenseNet201) can perform better than the independent model.

J. X. Yu et al. [21] proposed an ensemble average of Convolutional Neural Networks that combines several pre-trained CNN models considering the importance of facial expression. The proposed ensemble model consists of training each pre-trained CNN model through a classification layer first combined with a multilayer perceptron. The newly formed model is fitted to the facial expression dataset. The predictions returned by all models are combined with the average model to determine the final class probability distribution. The ensemble average model of the proposed CNN models is evaluated on three facial expression datasets: FER-2013, improved CK+, and RAF-DB. Since the improved CK+ dataset is a small dataset, data augmentation was used to increase the data size and diversity. In addition, oversampling was adopted to solve the class imbalance problem in RAF-DB. The empirical results show that the proposed CNN ensemble average model outperforms the individual ensemble model in terms of test accuracies of 77.70%, 94.10%, and 87.50% on the improved CK+ and RAF-DB datasets in FER 2013, respectively.

The ensemble model proposed by R. Lawpanom et al. [22] was developed using a homogeneous ensemble convolutional neural network called HoE-CNN for future online learning. The HoE-CNN ensemble model was trained on the FER 2013 dataset, which contains seven basic classes (Angry, Disgust, Fear, Happy, Sad, Surprised, Neutral). The results reported by the researchers show that ensemble models of deep learning models perform better than a single deep learning model. The efficiency of the proposed model classification results and the transfer of the model application to online learning applications, considering the uneven datasets and multi-class classifications, achieved 75.51% accuracy on the FER2013 dataset.

3. Dataset Preparation

The field of emotion recognition has developed significantly in the last decade, with deep learning methods such as convolutional neural networks playing a significant role. Traditionally, this field has used datasets such as FER2013, CK+ [30]or AffectNet [31], which typically cover 6 or 7 basic emotions (anger, disgust, fear, happiness, sadness, surprise and neutral). However, the limitations of these datasets – such as the small number of classes, poor image quality or limited exposure to real-world conditions – have led to the need to develop new and larger datasets. To address the issues raised in this research work, we propose our FaceEmocDS dataset. The proposed dataset was developed by correcting the shortcomings of other similar datasets. Figure 3 shows examples of images from the proposed FaceEmocDS dataset. The proposed dataset is designed to be divided into eight main classes: “anger”, “contempt”, “disgust”, “fear”, “happy”, “neutral”, “sad” and “surprise”.

To create the dataset, 277923 images were extracted from the 10 most popular datasets, manually analyzed, and non-compliant images were removed. The FaceEmocDS dataset contains a total of 72,412 face images. The merging process addressed the shortcomings of previous datasets. The images in the dataset must be taken from publicly available images or video footage without violating the privacy of personal data. All images in our collection are collected from open sources on the Internet, social media. Because the images in our collection are collected from previously used images, found through Internet searches, and from film clips. **Image quality** – the image quality must be sharp and have sufficient contrast to detect fine features. The popular dataset FER-2013 has an image size of 48x48 [32], The size of the images in the RAF-DB dataset is 100x100. In general, the images vary in size when they are collected. Most of the images in the old dataset are of poor quality (Figure 4), and such images were identified and removed. There is a problem with image size uniformity, which is that when resizing an image with the usual resize() functions, the image quality is degraded, so the ESPCN artificial intelligence model was used to increase the size of the image.

3.1. ESPCN (Efficient Sub-Pixel Convolutional Neural Network)

ESPCN is an efficient model for high-quality image upscaling, which implements super-resolution processing using sub-pixel convolutions. This model was introduced by Chao Dong et al. in 2016. ESPCN has a simple structure and uses computational resources economically [33]. The main idea of the ESPCN model is to upscale the image using traditional methods, such as bicubic interpolation, and then train neural networks based on low-resolution, and finally generate a high-resolution image through a sub-pixel layer. Figure 1 shows the structure of the ESPCN network [34]. This improves the efficiency of the model.

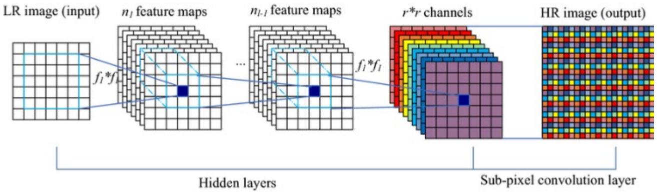

Figure 1. ESPCN network structure.

The loss function of the ESPCN network is:

$$L(W_{1:L}, b_{1:L}) = \frac{1}{r^2HW} \sum_{x=1}^{rH} \sum_{x=1}^{rW} (I_{x,y}^{HR} - f_{x,y}^{L}(I^{LR}))^2, \quad (1)$$

Where I($^{\mathrm{HR}}$) represents each original image in the dataset; I($^{\mathrm{LR}}$) represents each downsampled LR image; r represents the upscaling factor; H represents the height value of the image; W represents the width value of the image, W($_{1:L}$) represents all the network weights to be learned, and b($_{1:L}$) represents all the possible learning values. In Figure 2, we can see the image size scaling in the usual case and the scaling using the ESPCN model.

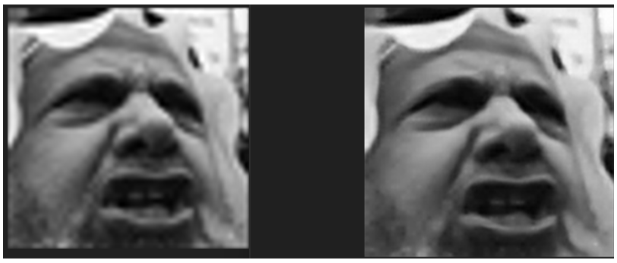

Figure 2. Comparison of resizing an image using a simple resize() function versus the ESPCN model.

In some cases, you may need to reduce the size of an image, in which case you can use the standard resize function.

3.2. Diversity of facial images

The diversity of images in the set is very important. In this case, the artificial intelligence model being trained will be adaptable to many conditions and situations. The diversity of images in the dataset should be ensured by demographic diversity, i.e., age-related differences, male and female, different nationalities and skin colors (for example, Asians, Africans, Europeans). In addition, it is important to ensure differences in facial position and background images. These parameters were taken into account when forming the FaceEmocDS set that we propose. However, one of the shortcomings of very popular datasets such as FER-2013, AffectNET, NHFI , JAFFE, RAF-DB, CK+ is the repetition of homogeneous images. We wrote a **hash_compare algorithm** during the problem set generation and used it to remove all unnecessary duplicates from the dataset. To write hash_compare(), we first read all the files in the directory. Then we calculate the hash value of each of them and compare it with the hash value of the previous image. This is a very effective practice, for example, when examining images in the “disgust” class in the AfectNET dataset, 54 identical images were identified. Samples in the datasets that did not show the full face image were also removed, so the accuracy of distinguishing individual facial features was also improved.

3.3. Balanced data

The number of data in each category (class) of the dataset should be equal or at least close to balance. The distribution of classes in the proposed FaceEmocDS dataset is: anger: 9390 images, contempt: 5002 images, disgust: 9405 images, fear: 9851 images, happy: 9554 images, neutral: 9722 images, sad: 10093 images, surprise: 9395 images. Since 277K images **were analyzed manually in this process,** special attention was paid to the balance of the number of images in the analysis classes by class, whereas, for example, in the FER-2013 dataset, there are 4953 face images in the “anger” class, but 547 samples in the “disgust” class. Such shortcomings have been sufficiently eliminated in the proposed FaceEmocDS dataset.

3.4. Accurately labelled data

Each image must be properly labelled. If facial location or facial structure is required, bounding box (face coordinates) or landmark (eye, nose, lip points) data must be available. In general, emotion recognition is a common instinctive process for humans, but it can be a rather complex problem even for very intelligent models. Therefore, it is very important for the model to be trained to correctly classify the datasets. During the development of the FaceEmocDS dataset, we performed the labeling process by analyzing each image. During the analysis, it was found that some facial images were incorrectly labeled, for example, an angry face was placed in another class, and such shortcomings were eliminated.

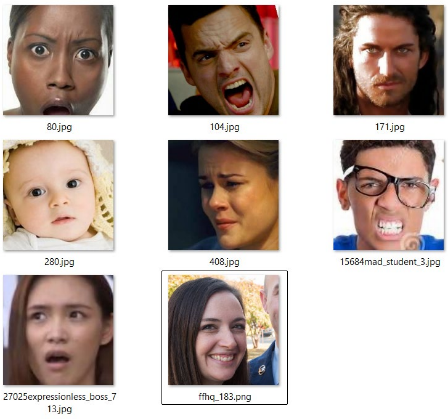

Figure 3. A sample of facial images from the FaceEmocDS dataset.

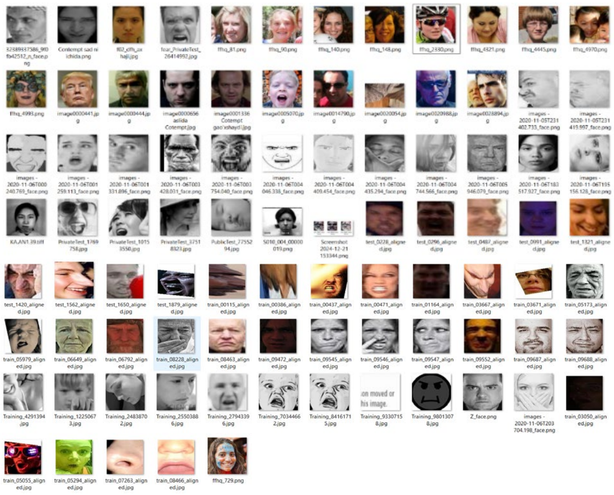

Figure 4. Examples of facial images that do not meet the requirements of the dataset.

The FaceEmocDS dataset was developed to provide a high-quality, diverse, and balanced database for research on facial expression and emotion recognition. This dataset was created by combining various popular datasets, eliminating their shortcomings, and improving them with advanced technologies. This dataset serves as a strong foundation for building accurate and efficient models in the field of facial expression recognition and emotion classification. FaceEmocDS not only corrected the shortcomings of existing datasets, but also provided the opportunity to use it as a reliable and open source for scientific research. This dataset is an important resource that can be applied in the fields of artificial intelligence and deep learning. Figure 6 shows a comparative analysis of the FaceEmocDS dataset with other popular datasets.

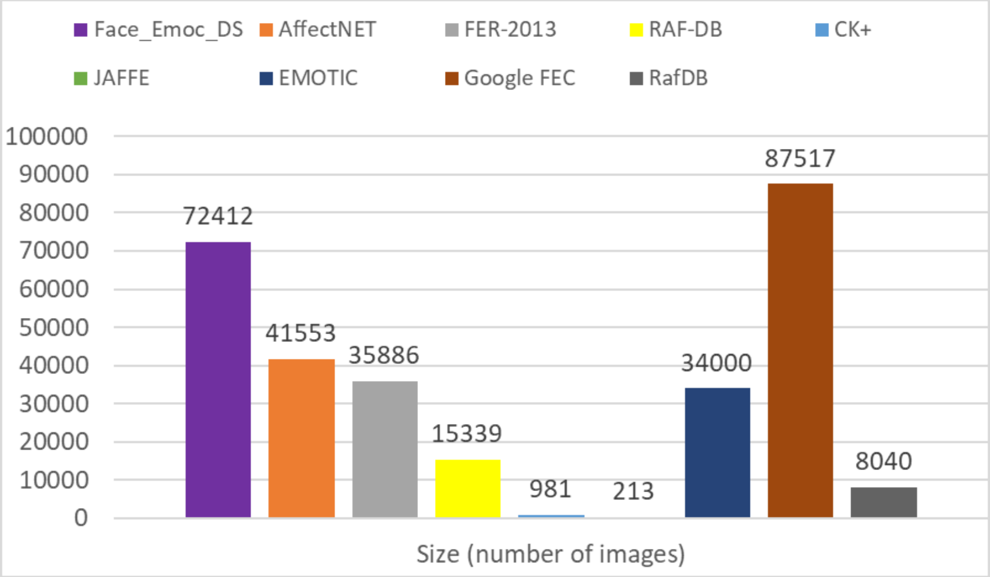

Figure 5. Comparative analysis of the proposed FaceEmocDS dataset and other datasets.

4. Materials and methods

An ensemble model is a meta-classifier that combines the predictions of multiple primary classifiers to form a more accurate final prediction. Ensemble classifiers are used in a variety of fields, including computer vision, speech recognition, and machine learning. Ensemble classifiers generally outperform single-model classifiers [23]. Ensemble models were originally used in machine learning algorithms [24]. They are now also widely used in deep learning model integration. There are five main ways to train ensemble models:

Bagging (Bootstrap Aggregating): Various deep learning models are trained on random subsets of data (bootstrap samples). Each model runs independently, and the results are pooled together to get an average or maximum value. For example, multiple Convolutional Neural Networks (CNNs) are trained on different random datasets and used for image classification [25].

Boosting – Deep learning models are trained sequentially, with each subsequent model focusing on the errors of the previous model. Boosting is less commonly used because deep networks require a lot of resources to train. However, approaches similar to Gradient Boosting can be adapted to deep learning models [26].

Stacking – In the first level, different deep learning models (e.g. CNN, LSTM, Transformer) are trained. Their predictions are used as new data, and a second level meta-model (e.g. simple neural network or logistic regression) is trained. In natural language processing, predictions from BERT, RoBERTa, and XLNet models are combined through stacking [27].

Bayesian Model Averaging (BMA) – multiple deep learning models with the same architecture (but with different initial weights or training parameters) are trained and their results are averaged [28]. This technique is common in deep learning because neural networks can produce different results depending on random initial weights.

Dropout-based Ensembles – In deep learning, dropout (randomly deleting neurons) is used to create an ensemble effect during training. Each dropout iteration creates a different "subnetwork". The results of multiple dropout iterations are combined during testing [29].

The ensemble model we propose is developed by combining previously obtained results. In this case, the VGG19, ResNet50 and DenseNet121 models, previously trained on the specially designed FaceEmocDS dataset, were trained in a classified form for eight emotions and achieved 74.61% accuracy in epoch 29, 75.92% in epoch 30 and 76.69% in epoch 22, respectively. Each independently produces its own result (emotion probabilities) based on the input images. These results are then combined into a single common vector. A Dropout operation is applied to this combined vector, which helps to prevent the model from overfitting. Then the final layer – a simple linear layer – processes this common vector and outputs the final prediction (which emotion it belongs to).

This approach is called stacked generalization (or stacking) because it combines the output of several basic models and uses a separate "metamodel" to make a final decision. This method often allows for correct classification of samples that individual models would not be able to correctly identify independently, because each model learns from different aspects of the data [27].

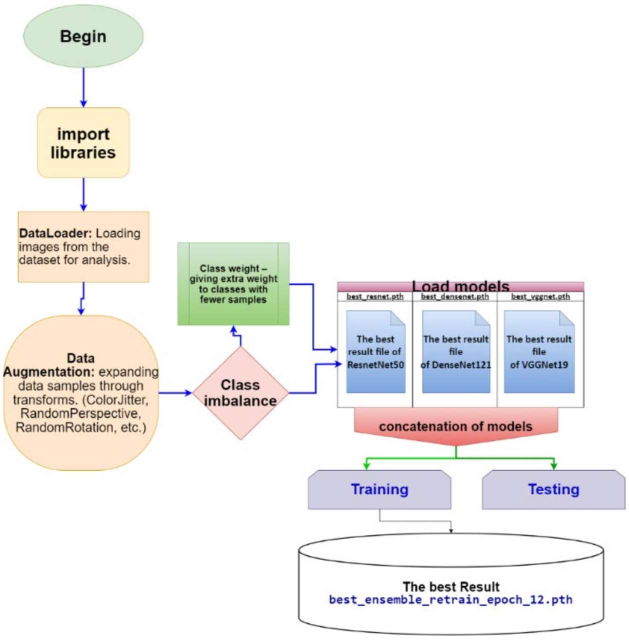
Begin
import libraries
DataLoader: Loading images from the dataset for analysis.
Data Augmentation: expanding data samples through transforms. (ColorJitter, RandomPerspective, RandomRotation, etc.)
Class imbalance
Class weight – giving extra weight to classes with fewer samples
Load models
best_resnet.pth
best_densenet.pth
best_vggnet.pth
The best result file of ResnetNet50
The best result file of DenseNet121
The best result file of VGGNet19
concatenation of models
Training
Testing
The best Result
best_ensemble_retrain_epoch_12.pth

Figure 6. The proposed combined ensemble model algorithm.

4.1. Preparing data for the ensemble model

Transformations - the transformation part of the process defines the process of preparing images for training and testing, and increasing the number of images.

(i) transforms.Lambda: Checks if the images are in RGB format and converts them to RGB if they are in another format (e.g. grayscale). This is necessary so that the model requires the same input format.

(ii) transforms.Resize((224, 224)): Resizes images to 224x224, which is the default input size for VGG, ResNet, and DenseNet models.

(iii) transforms.ColorJitter: Randomly jitters the brightness, contrast, saturation, and hue (each in a range of 0.2). This augmentation method increases the model's resilience to changes in lighting conditions.

(iv) transforms.RandomResizedCrop: Randomly crops and resizes the image (scale=0.8-1.0). This improves the model's ability to focus on different parts of the face.

(v) transforms.RandomPerspective: Distorts the image with a random perspective shift of 0.2 degrees (with a 50% probability). This simulates changes in the angles of the face.

(vi) transforms.RandomHorizontalFlip: Flips the image horizontally (50% probability). This helps to study the symmetrical features of the face.

(vii) transforms.RandomRotation(10): Randomly rotates the image within ±10 degrees, ensuring that it adapts to small angular changes of the face.

(viii) transforms.ToTensor(): Converts the image to PyTorch tensor format (normalizes values 0-255 to the range 0-1).

(ix) transforms.Normalize: Normalizes the image to the ImageNet mean and standard deviation values. This is standard practice for pre-trained models.

Test transform (test_transform) : Unlike training, the test transform only includes the processes of homogenization, tensor conversion, and normalization. Augmentation methods are not used because the original state of the images must be preserved during the testing process.

The training transform uses a wide range of augmentation techniques to improve the generalization ability of the model, while the testing transform provides a stable and standardized input. These two approaches are consistent with the goals of the training and testing processes.

4.2. Model loading

Involves fitting pre-trained VGG19, ResNet50, and DenseNet121 models to *the FaceEmocDS dataset:*

- **model_class(pretrained=False)** : The model is loaded without pretraining on ImageNet, as we are using our own stored weights (*vgg19_epoch_30.pth* , etc.).
- **VGG adaptation** : model.classifier[6] = nn.Linear(4096, 8) – The last classification layer of VGG19 is changed from 1000 classes to 8 classes. This layer projects a 4096-dimensional feature vector into 8 emotion classes.
- **ResNet adaptation** : model.fc = nn.Sequential(...) – The last layer of ResNet50 is replaced with a two-stage network: an intermediate layer with 2048 to 512 neurons (with ReLU) and a final layer with 512 to 8 classes. This change reduces memory requirements and increases flexibility.
- **DenseNet adaptation** : model.classifier = nn.Linear(1024, 8) – The last layer of DenseNet121 classifies a 1024-dimensional vector into 8 classes.
- **torch.load and load_state_dict** : Loads saved model files, while strict=False allows you to ignore incompatible layers (e.g. old ImageNet outputs).
- **to(device)** : The model is placed on the GPU.
- This function ensures that the modules are adapted and the stored weights are loaded accordingly. Appropriate changes are made to the architecture of each model to ensure full adaptation to the 8-class task.

4.3. Ensemble structure

The structure of the ensemble model is based on combining the outputs of three models:

__init__ : Takes VGG, ResNet, and DenseNet models as internal components. Dropout(0.3) is added to reduce overfitting. nn.Linear(24, 8) combines the outputs of the three models into 8 classes (3 x 8 = 24).

forward : Each model outputs an 8-dimensional vector from the input image. torch.cat merges these vectors into a 24-dimensional vector (by dim=1, i.e. the axis after the batch size). The dropout layer randomly removes 30% of the neurons. The final linear layer transforms the 24-dimensional vector into an 8-class output.

The ensemble model uses a simple but effective approach: it combines the features of three models to make a final decision. The addition of dropout serves to increase the generalization ability of the model.

5. Results

The results were analyzed based on the final ensemble accuracy of the model (Final Ensemble Accuracy: 0.8566) and the metrics in the classification report – precision, recall, F1-score and support – and the model performance across epochs was discussed as a general trend. The test dataset contains 14,483 images and was evaluated based on the best result from the training process.

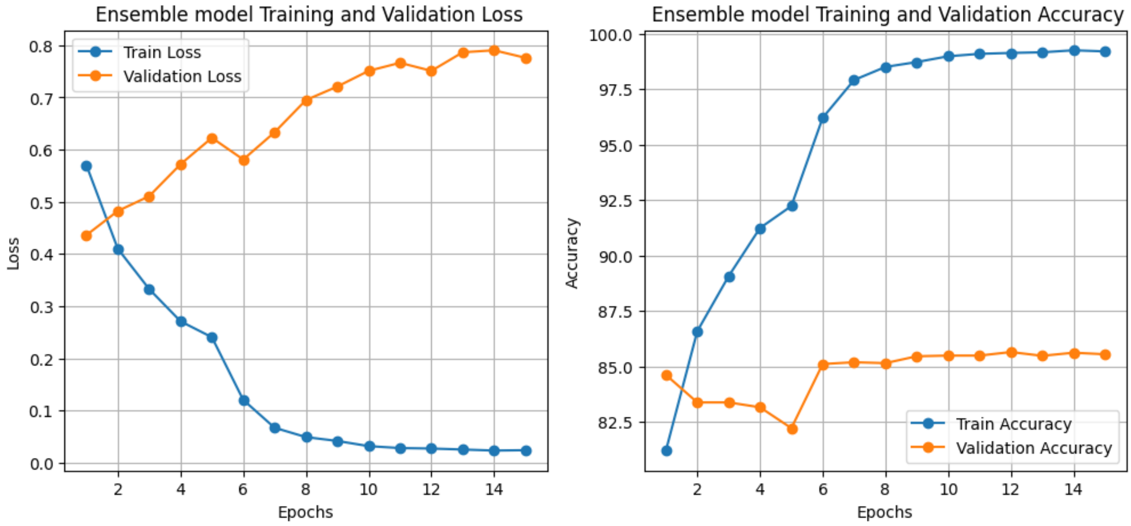

Figure 7. Training results of the proposed Ensemble model.

The overall accuracy of the model is 85.66%, which is considered a high result for a complex dataset with 8 classes. The Classification report provided the following metrics for each class:

- **Anger** : Precision: 0.83, Recall: 0.85, F1-score: 0.84, Support: 1870
- **Contempt** : Precision: 0.70, Recall: 0.76, F1-score: 0.73, Support: 1000
- **Disgust** : Precision: 0.87, Recall: 0.84, F1-score: 0.85, Support: 1884
- **Fear** : Precision: 0.81, Recall: 0.84, F1-score: 0.83, Support: 1989
- **Happy** : Precision: 0.97, Recall: 0.96, F1-score: 0.97, Support: 1844
- **Neutral** : Precision: 0.85, Recall: 0.86, F1-score: 0.86, Support: 1937
- **Sad** : Precision: 0.88, Recall: 0.85, F1-score: 0.86, Support: 2047
- **Surprise** : Precision: 0.88, Recall: 0.84, F1-score: 0.86, Support: 1912

Accuracy: 0.8566, **Macro Avg**: Precision: 0.85, Recall: 0.85, F1-score: 0.85, **Weighted Avg**: Precision: 0.86, Recall: 0.86, F1-score: 0.86, **Support**: 14,483

The overall accuracy shows that approximately 12,408 images out of 14,483 images were correctly classified. The Macro Avg F1-score (0.85) reflects an even balance across classes, while the Weighted Avg F1-score (0.86) reflects the effect of the number of images. The “Happy” class had the highest result (F1-score: 0.97), while the “contempt” class had the lowest result (F1-score: 0.73).

5.1. Analysis by epochs

The model was trained for 15 epochs, and the training (Train Loss, Train Acc) and validation (Val Loss, Val Acc) results were recorded for each epoch. This analysis helps to assess the model's training dynamics, signs of overfitting, and the process of reaching optimal results.

Epoch 1: Train Loss: 0.5701, Train Acc: 81.22%, Val Loss: 0.4363, Val Acc: 84.63%, Time: 2124.00s – In the first epoch, the training accuracy was 81.22%, and the model began to learn the main features of the dataset. Validation accuracy reached 84.63%, which is considered a high initial result for an 8-class dataset. Val Loss 0.4363 is lower than Train Loss, confirming that the model generalizes well on the validation data. This result was saved as the initial best model and indicates a successful start of training.

Epoch 6: Train Loss: 0.1205, Train Acc: 96.22%, Val Loss: 0.5811, Val Acc: 85.12%, Time: 1870.34s – In the sixth epoch, the training accuracy increased sharply to 96.22%, while the Train Loss decreased to 0.1205, indicating that the model had learned the training data to a high degree. The validation accuracy increases to 85.12%, an improvement from Epoch 1 (84.63%). The Val Loss increased to 0.5811, which is higher than the 0.4363 in Epoch 1, but this result was maintained as the best model due to the increase in Val Acc. The widening of the gap between training and validation shows early signs of overfitting, but the improvement in the validation result confirms that the overall performance of the model has increased.

Epoch 7: Train Loss: 0.0671, Train Acc: 97.92%, Val Loss: 0.6333, Val Acc: 85.20%, Time: 1901.16s – In the seventh epoch, the training accuracy reached 97.92%, with a significant decrease in Train Loss to 0.0671, indicating that the model was almost perfectly fitted to the training data. The validation accuracy increased to 85.20%, while Val Loss increased to 0.6333. This result was retained as the best model due to a slight improvement in Val Acc (from 85.12% to 85.20%). However, the increase in Val Loss (from 0.5811 to 0.6333) indicates increased overfitting, as the model increased the error rate on the validation data.

Epoch 9: Train Loss: 0.0418, Train Acc: 98.73%, Val Loss: 0.7209, Val Acc: 85.47%, Time: 1822.13s – In the ninth epoch, the training accuracy increased to 98.73%, while the Train Loss decreased to 0.0418, indicating that the model learned the training data better. The validation accuracy reached 85.47%, while the Val Loss increased to 0.7209. The increase in Val Acc (from 85.20% to 85.47%) maintained this result as the best model . The significant increase in Val Loss (from 0.6333 to 0.7209) indicates that overfitting has increased further, but the improvement in validation accuracy has maintained the overall reliability of the model.

Epoch 10: Train Loss: 0.0319, Train Acc: 98.99%, Val Loss: 0.7510, Val Acc: 85.50%, Time: 1822.09s – In the tenth epoch, the training accuracy increased to 98.99%, Train Loss decreased to 0.0319. The validation accuracy reached 85.50%, Val Loss increased to 0.7510. The small increase in Val Acc (from 85.47% to 85.50%) kept this result as the best model. The increase in Val Loss (from 0.7209 to 0.7510) indicates that overfitting is still occurring, but the steady increase in validation accuracy confirms that the model is approaching its optimal point.

Epoch 12: Train Loss: 0.0273, Train Acc: 99.14%, Val Loss: 0.7510, Val Acc: 85.66%, Time: 1821.73s – At the twelfth epoch, the training accuracy reached 99.14%, while the Train Loss dropped to 0.0273, indicating that the model classified the training data almost perfectly. The validation accuracy increased to 85.66%, which was the highest result, while the Val Loss remained stable at 0.7510. This result was retained as the best model, and the final accuracy in the test (85.66%) is assumed to be obtained from this model. Despite the high Val Loss, the fact that Val Acc reached its maximum level indicates the best balance of the model’s generalization ability.

The analysis by epochs sheds light on the training process of the model. The training accuracy increased from 81.22% to 99.26%, while the Train Loss decreased from 0.5701 to 0.0234, indicating that the model learned the training data almost perfectly. The validation accuracy increased from 82.23% to 85.66%, with the best result recorded at Epoch 12 (Val Acc: 85.66%, Val Loss: 0.7510). However, the increase in Val Loss from 0.4363 to 0.7904 indicated overfitting, especially after Epoch 5 (Val Acc: 82.23%, Val Loss: 0.6227). After Epoch 12, Val Acc remained stable (85.49%-85.66%), while Val Loss remained high (0.7510-0.7904), indicating the limits of the model's generalization ability. The 85.66% accuracy in the test is consistent with the Epoch 12 result, confirming the reliability of the best model.

5.2. Analysis by class

Anger – Precision: 0.83, Recall: 0.85, F1-score: 0.84, Support: 1870. Recall 0.85 indicates that approximately 1590 images out of 1870 images of “anger” were correctly identified. Precision 0.83 means that 83% of the predicted images are correct. F1-score 0.84 indicates stable classification of the class.

Contempt – Precision: 0.70, Recall: 0.76, F1-score: 0.73, Support: 1000. Recall 0.76 indicates that out of 1000 images of “contempt” approximately 760 images were correctly identified. Precision 0.70 is low, meaning that 30% of the predictions are incorrect. F1-score 0.73 indicates the weak point of the model as the lowest result.

Disgust – Precision: 0.87, Recall: 0.84, F1-score: 0.85, Support: 1884. Recall 0.84 indicates that approximately 1583 images out of 1884 images of “disgust” were correctly identified. Precision 0.87 is high, and F1-score 0.85 confirms good classification of the class.

Fear – Precision : 0.81, Recall: 0.84, F1-score: 0.83, Support: 1989. Recall 0.84 indicates that “fear” was correctly identified in approximately 1671 images out of 1989 images. Precision 0.81 is average, F1-score 0.83 indicates that the class is stable but not optimal.

Happy – Precision: 0.97, Recall: 0.96, F1-score: 0.97, Support: 1844. Recall 0.96 indicates that approximately 1770 images out of 1844 images of “happy” were correctly identified. Precision 0.97 is high, F1-score 0.97 indicates the most successful class of the model.

Neutral – Precision: 0.85, Recall: 0.86, F1-score: 0.86, Support: 1937. Recall 0.86 indicates that approximately 1666 images out of 1937 images of “neutral” were correctly identified. Precision 0.85, F1-score 0.86 confirms stable classification of the class.

Sad – Precision: 0.88, Recall: 0.85, F1-score: 0.86, Support: 2047. Recall 0.85 indicates that approximately 1740 images out of 2047 images of “sad” were correctly identified. Precision 0.88 is high, F1-score 0.86 indicates good classification of the class.

Surprise – Precision: 0.88, Recall: 0.84, F1-score: 0.86, Support: 1912. Recall 0.84 indicates that approximately 1606 images out of 1912 images of “surprise” were correctly identified. Precision 0.88, F1-score 0.86 confirms the stable result of the class.

6. Discussion

The overall accuracy of the model is 85.66%, indicating that approximately 12,408 images out of 14,483 are correctly classified. The Macro Avg F1-score (0.85) shows a balance across classes, but the low result of "contempt" (F1-score: 0.73) reduces this indicator. The Weighted Avg F1-score (0.86) is improved due to the higher classes (sad: 2047, fear: 1989) considering the number of images. While "Happy" (F1-score: 0.97) is the highest result, "contempt" (F1-score: 0.73) stands out as the weakest point.

Analysis by epochs shows that the model learned clear expressions ("happy", "sad") quickly in the early stage, improved on classes such as "disgust" and "neutral" in the middle stage, and improved on more complex classes such as "contempt" and "fear" in the final stage, but did not achieve complete success. The high performance of "Happy" is due to its clear features and good distribution in the dataset (1844 images). The "Sad" and "surprise" classes also showed high performance, which confirms that their features were well learned by the model. The low performance of "Contempt" may be due to its small number of images in the dataset (1000) and its fine-grained representation.

In this study, the proposed ensemble model and its training results on the FaceEmocDS dataset were compared with the results obtained in other studies. Table 1 presents this comparative analysis.

Table 1. Comparison of the results obtained with the results of other studies

Research	Data set	Combined models for ensemble	Accuracy	Note
E.G. Moung et al. (2024) [2]	FER2013	Custom CNN, ResNet50, InceptionV3	72.3%	Lower result than the model we proposed
V. Shtino et al. (2024) [3].	AffectNet	Inception V3, ResNet50, SPP-net	85.2%	Lower result than the model we proposed
J.X. Chang et al. (2023) [4]	CK+, FER2013, and JAFFE	VGG-19, VGGFace, ViT-B/16, and ViT-B/32	100%, 76.30%, 100%	Very good result on a small dataset, but inferior to our proposed model on the relatively larger FER2013
A. Das et al. [5]	FER2013	DenseNet121, DenseNet169, DenseNet201	74.16	It performed well for the FER2013 dataset, but accuracy is lower for larger datasets.
Yu, Jing Xuan et al. (2021) [6]	FER-2013, improved CK+ and RAF-DB	VGG16, VGG19, ResNet50, ResNet101	77.70%, 94.10% 87.50%	The RAF-DB dataset is smaller than the Face_Emoc_DS dataset.
M. K. Naveen et al. (M and S. 2023).	FER-2013	Facenet, Facenet2018, VGG16, Resnet-50, Senet-50, Arcface, Openface	74.67%	It performed well for the FER2013 dataset, but accuracy is lower for larger datasets.
Lawpanom R, et al. [7]	FER2013	A homogeneous ensemble convolutional neural network called HoE-CNN	75.51%	It performed well for the FER2013 dataset, but accuracy is lower for larger datasets.
Proposed ensemble model	Face_Emoc_DS	VGGNet19, ResNet50, DenseNet121	85.66%	

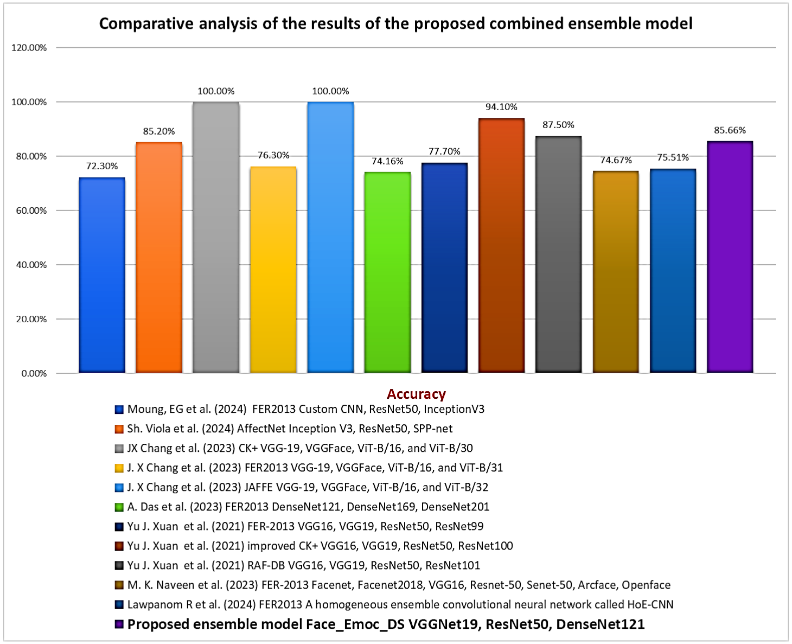

Figure 8. Comparative analysis of the results of the proposed combined ensemble model.

7. Conclusion

When we compare the results obtained in the studies with the results of the model we proposed, demonstrated a number of important advantages in the field of emotion recognition from facial expressions. These advantages are related to the characteristics of the dataset, the methodological approach, and the practical significance of the results obtained, highlighting the effectiveness of the proposed solution in real-world conditions.

First of all, the large size (72,412 images) and high-quality images (224x224 RGB format) of *the FaceEmocDS* dataset distinguish it from popular datasets such as FER2013 (35,887 images, 48x48 grayscale), SFEW (1,246 images), and even AffectNet (450,000 images, but poor classification quality). For example, while the low-resolution images in FER2013 made it difficult to detect subtle expressions (such as "Disgust" or "Fear"), the high-resolution color images of FaceEmocDS allowed us to more clearly distinguish important facial features – eye opening, mouth opening, eyebrow position. This feature provided a significant advantage in the classes "Happy" (F1 = 0.97) and "Disgust" (F1 = 0.85) compared to FER2013 (Happy ≈ 0.90, Disgust ≈ 0.58) and AffectNet (Happy ≈ 0.85-0.90, Disgust ≈ 0.55). At the same time, the dataset was collected in real-world conditions, including different lighting conditions, facial angles, and demographic diversity, which makes it more practical than small datasets collected in laboratory conditions such as CK+ (981 images) and JAFFE (213 images) [36]. Google Facial Expression Comparison (FEC) is larger than the proposed dataset, but it contains 30 different emotions and many emotions that are very close to each other. This makes it difficult for the model to learn.

The second advantage is the methodological superiority of the ensemble model. The combination of VGG19, ResNet50 and DenseNet121 models allows us to take advantage of the strengths of each architecture: the ability to extract detailed features of VGG19, the stable learning process through residual connections of ResNet50 and the efficiency of feature reuse of DenseNet121. This approach improves the results of individual models (ResNet-50 – 75.92%, VGGNet-19 – 74.50%, DenseNet-121– 76.10%, respectively) to 85.66%. This result shows the superiority of other research results using the ensemble model method.

The third advantage is class diversity and balance. *FaceEmocDS* covers 8 classes (Anger, Contempt, Disgust, Fear, Happy, Neutral, Sad, Surprise), which is more than studies such as FER2013 (7 classes), RAF-DB (7 classes) and SFEW (7 classes). Although the addition of the "Contempt" class complicates the task, the F1-score achieved in this class (0.73) is significantly higher than the results in AffectNet (0.50-0.60), which is of practical importance for detecting subtle expressions in real life. Class balance (image range 5002-10,093) was achieved using *class weights to reduce class imbalance* , which showed significant improvement in classes such as "Disgust" (0.85) and "Fear" (0.83) compared to FER2013 (Disgust ≈ 0.58, Fear ≈ 0.65) and AffectNet (Disgust ≈ 0.55, Fear ≈ 0.60).

Finally, the improved performance of FaceEmocDS in real-world conditions (85.66%) makes it possible to actively use it in practical applications - in areas such as human-computer interaction systems, psychological analysis and marketing. For example, while high performance on controlled datasets such as CK+ (95-99.26%) and JAFFE (92-98%) is limited to laboratory tests, our proposed ensemble model is closer to real-world results due to its testing on a large dataset.

References

[1] G. Perveen, S. F. Ali, J. Ahmad, S. Shahab, M. Adnan and Mohd Anjum, "Multi-Stream Deep Convolution Neural Network With Ensemble Learning for Facial Micro-Expression Recognition," IEEE Access, vol. 11, pp. 118474-118489, 2023.

[2] M. Almubarak and F. Alsulaiman, "An Ensemble Learning Approach for Facial Emotion Recognition Based on Deep Learning Techniques," Electronics, vol. 14, p. 3415, 2025.

[3] G. Zhou, Y. Xie and W. Tian., "Multi loss-based feature fusion and top two voting ensemble decision strategy for facial expression recognition in the wild.," 2023. [Online]. Available: arXiv:2311.03478.

[4] E. Younis, S. Zaki, E. Kanjo and E. Houssein, "Evaluating Ensemble Learning Methods for Multi-Modal Emotion Recognition Using Sensor Data Fusion.," Sensor, vol. 22, p. 5611, 2022.

[5] A. Renda, M. Barsacchi, A. Bechini and F. Marcelloni, "Comparing ensemble strategies for deep learning: An application to facial expression recognition," Expert Systems with Applications, vol. 136, pp. 1-11, 2019.

[6] D. Pandit and S. Jadhav, "Fusion of ensemble technique with CNN model to equilibrate prediction of face emotions in real-time," Traitement du Signal, vol. 42, no. 1, pp. 519-530, 2024.

[7] M. Ganaie, M. Hu, A. Malik, M. Tanveer and P.N. Suganthan, "Ensemble deep learning: A review," Engineering Applications of Artificial Intelligence, vol. 115, p. 105151, 2022.

[8] E. A. Sağbaş, A. Uğur and S. Korukoğlu, "Performance Evaluation of Ensemble Learning Methods for Facial Expression Recognition,," in Innovations in Intelligent Systems and Applications Conference (ASYU), Izmir, Turkey, 2019.

[9] J. Y. Choi and B. Lee, "Combining Deep Convolutional Neural Networks With Stochastic Ensemble Weight Optimization for Facial Expression Recognition in the Wild," IEEE Transactions on Multimedia, vol. 25, pp. 100-111, 2023.

[10] A. M. Esfar-E-Alam, M. Hossain, M. Gomes, R. Islam and R. Raihana, "Multimodal Emotion Recognition Using Heterogeneous Ensemble Techniques,," in International Conference on Computer and Information Technology (ICCIT), Cox's Bazar, Bangladesh, 2022.

[11] W. Li, M. Luo, P. Zhang and W. Huang, "A Novel Multi-Feature Joint Learning Ensemble Framework for Multi-Label Facial Expression Recognition," IEEE Access, vol. 9, pp. 119766-119777, 2021.

[12] H. D. Nguyen, S. H. Kim, G. S. Lee, H. J. Yang, I. S. Na and S. H. Kim, "Facial Expression Recognition Using a Temporal Ensemble of Multi-Level Convolutional Neural Networks," IEEE Transactions on Affective Computing, vol. 13, no. 1, pp. 226-237, 2019.

[13] E. S. Smitha, S. Sendhilkumar and G. S. Mahalakshmi, "Ensemble Convolution Neural Network for Robust Video Emotion Recognition Using Deep Semantics," Scientific Programming, p. 21, 2023.

[14] M. Muhajir, K. Muchtar, M. Oktiana and A. Bintang, "Students' emotion classification system through an ensemble approach," SINERGI, vol. 28, no. 2, pp. 413-424, 2024.

[15] S. Gupta, P. Kumar and RajKumar Tekchandani, "EDFA: Ensemble deep CNN for assessing student's cognitive state in adaptive online learning environments," International Journal of Cognitive Computing in Engineering, vol. 4, pp. 373-387, 2023.

[16] H. Dương, S. Yeom, G.-S. Lee, H.-J. Yang, I. Na and S. Kim, "Facial Emotion Recognition Using an Ensemble of Multi-Level Convolutional Neural Networks.," International Journal of Pattern Recognition and Artificial Intelligence., vol. 33, 2019.

[17] E. G. Moung, C. C. Wooi, M. M. Sufian, J. A. Dargham and J. Khoo, "A Robust Ensemble Approach to Face Expression Recognition and Image Sentiment Analysis.," in Internet of Things and Artificial Intelligence for Smart Environments., Singapore, Springer, 2024, p. 203–223.

[18] V. Shtino and M. Muca, "Improving Facial Expression Classification through Ensemble Deep Learning Models. Nanotechnology Perceptions," Nanotechnology and the Applications in Engineering and Emerging Technologies, vol. 20, 2024.

[19] J. X. Chang, C. P. Lee, K. M. Lim and J. Y. Lim, "Facial Expression Recognition with Machine Learning," in 11th International Conference on Information and Communication Technology (ICoICT), Malaysia, 2023.

[20] A. Das, M. S. Jalal, A. Bari and M. Huda, "Facial Emotion Recognition by Ensemble-DenseNet Networks," in IEEE 11th Region 10 Humanitarian Technology Conference (R10-HTC), Rajkot, India,, 2023.

[21] J. X. Yu, K. M. Lim and C. P. Lee, "MoVE-CNNs: Model aVeraging Ensemble of Convolutional Neural Networks for Facial Expression Recognition.," IAENG International Journal of Computer Science,, vol. 48, no. 3, pp. 1-5, 2021.

[22] R. Lawpanom, W. Songpan and J. Kaewyotha, "Advancing Facial Expression Recognition in Online Learning Education Using a Homogeneous Ensemble Convolutional Neural Network Approach.," Applied Sciences, vol. 14, no. 3, p. 1156, 2024.

[23] M. Ashraf, M. Zaman and M. Ahmed, "An intelligent prediction system for educational data mining based on ensemble and filtering approaches.," Procedia Comput. Sci., vol. 167, pp. 1471-1483, 2020.

[24] scikit-learn.org, Available: https://scikit-learn.org/stable/modules/ensemble.html.

[25] F. H and S. Weijs, "Entropy Ensemble Filter: A Modified Bootstrap Aggregating (Bagging) Procedure to Improve Efficiency in Ensemble Model Simulation.," Entropy, vol. 19, no. 10, p. 520, 2017.

[26] N. Mungoli, "Adaptive ensemble learning: Boosting model performance through intelligent feature fusion in deep neural networks.," arXiv preprint, 2023.

[27] J. Brownlee, "https://machinelearningmastery.com/stacking-ensemble-for-deep-learning-neural-networks/," machinelearningmastery.com, August 2020. [Online]. Available: https://machinelearningmastery.com/stacking-ensemble-for-deep-learning-neural-networks/ .

[28] J. Montgomery, F. Hollenbach, M. Ward and R. Alvarez, "Improving Predictions Using Ensemble Bayesian Model Averaging.," Political Analysis., pp. 271-291, 2012.

[29] K. Hara, D. Saitoh and H. Shouno, "Analysis of Dropout Learning Regarded as Ensemble Learning," in Artificial Neural Networks and Machine Learning – ICANN 2016. ICANN 2016, 2019.

[30] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar and I. Matthews, "The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression," in IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops, San Francisco, 2010.

[31] A. Mollahosseini, B. Hasani and M. H. Mahoor, "Affectnet: A database for facial expression, valence, and arousal computing in the wild.," IEEE Transactions on Affective Computing, vol. 10, no. 1, pp. 18-31, 2017.

[32] Y. Khaireddin and Z. Chen, "Facial emotion recognition: State of the art performance on FER2013.," arxiv.org, 2021. [Online]. Available: https://arxiv.org/pdf/2105.03588.

[33] C. Dong, C. C. Loy and X. Tang, "Accelerating the Super-Resolution Convolutional Neural Network.," in In European conference on computer vision , 2016.

[34] H. Ruan, Z. Tan, L. Chen, W. Wan and J. Cao, "Efficient sub-pixel convolutional neural network for terahertz image super-resolution.," Optics letters, vol. 47, no. 12, p. 3115–3118., 2022.

[35] M. N. Kmuar and S. G. Winster, "Emotion identification in human faces through ensemble of deep learning models," Journal of Intelligent & Fuzzy Systems, vol. 45, no. 6, pp. 9729-9752, 2023.

[36] M. J. Lyons, "Excavating AI" re-excavated: debunking a fallacious account of the JAFFE dataset," arXiv, 2021. [Online]. Available: https://arxiv.org/pdf/2107.13998.

[37] J. Brownlee, "Stacking Ensemble for Deep Learning Neural Networks in Python," 28 07 2020. [Online]. Available: https://machinelearningmastery.com/stacking-ensemble-for-deep-learning-neural-networks/.