Undergraduate Conference on Intelligent Computing and Systems (UCICS 2025)
26-27 February, 2025; Varendra University, Rajshahi, Bangladesh

Human Emotion Recognition Utilizing Transfer Learning

Most. Afia Sultana
Computer Science and Engineering
Varendra University
Rajshahi, Bangladesh
sultanaafia008@gmail.com

Md. Shahed Rahman
Computer Science and Engineering
Varendra University
Rajshahi, Bangladesh
rahmanmdshahed94@gmail.com

Nafisa Anjum
Computer Science and Engineering
Varendra University
Rajshahi, Bangladesh
n.anjum.bd@gmail.com

A.S.M Zubaer Haider
Computer Science and Engineering
Varendra University
Rajshahi, Bangladesh
zubaerhaider@gmail.com

Pallab Chowdhury
Computer Science and Engineering
Varendra University
Rajshahi, Bangladesh
chowdhurypall95@gmail.com

Md. Taufiq Khan
Computer Science and Engineering
Varendra University
Rajshahi, Bangladesh
khantaufiq2001@gmail.com

***Abstract*—Facial Emotion Recognition (FER) is a crucial area of research in the field of Computer Vision (CV) and Deep Learning (DL), aiming to identify and classify emotions such as happiness, sadness, anger, and surprise from facial expressions. This study focuses on developing a model for recognizing facial emotions using DL techniques along with Transfer Learning (TL). For this work, we have used CKPLUS dataset. We have used two pre-trained State-of-the-art (SOTA) Convolutional Neural Network (CNN) models, InceptionV3 and MobileNet, to transfer their learning in the task of feature extraction from our image data. For the classification task, we have used extracted features from these models both individually and fusing them together. For classifying these sets of features, we have used multiple classifiers, namely, Logistic Regression, a custom 1-dimensional (1D) CNN, and hard voting classifiers that ensemble the decisions of Linear Support Vector Machine (LSVM), Logistic Regression, perceptron, quadratically penalized SVM, and SVM with quadratically smooth loss. The fused features extracted from InceptionV3 and MobileNet achieve the highest performance when classified using the voting classifier, with a classification accuracy of 98.48%, a weighted precision of 99.53%, a recall of 98.57%, and an F1-score of 98.48%. This work indicates the ensemble learning technique outperforms individual models across the dataset, identifying and classifying human emotion from an image of facial expression more accurately.**

***Index Terms*—Facial emotion recognition, Image classification, Ensemble Learning.**

I. Introduction

Facial emotion recognition (FER) is an essential domain for machine learning (ML) and artificial intelligence (AI), where we infer human emotion based on the facial expressions of individuals. Embodied emotion is crucial for human communication, facilitating an intuitive channel through which intentions, thoughts, and feelings are communicated. FER systems have great potential in numerous areas, including healthcare, where they help diagnose psychological disorders, education, to provide a more personalized learning experience, and in customer service, where they provide sentiment analysis services [1] [2].

Handcrafted feature extraction methods, like Gabor filters and Local Binary Patterns (LBP) that characterized facial features based on texture and shape, were initially used for FER. Although successful under controlled conditions, these approaches did not perform well under variations of lighting, pose, and occlusions. The arrival of deep learning has had a transformative impact on FER, with convolutional neural networks (CNNs) learning features directly from data hierarchically and robustly. This has facilitated efficient advancement in the recognition of emotions, even in complex, real-world scenarios [3].

However, this requires overcoming hurdles to ensure effective functioning across heterogeneous demographics, emotional scales, and cultural contexts. To further improve the engagement of recognition systems in specific applications (like surveillance or interactive systems), real-time recognition of emotions should be achieved, while focusing on ensuring computational efficiency. Further, ethical considerations such as privacy and potential misuse of emotion recognition data need thorough contemplation to make sure that FER technologies are deployed responsibly [4] [5].

With the emergence of new technologies, recent studies have adopted a multimodal approach by integrating facial features with voice and physiological signals, resulting in increased accuracy. More recently, transformer-based models with attention mechanisms are being explored to capture global and context-dependent dependencies in facial expressions. With advancements in FER technology, there is the potential to close gaps in human-computer interaction and develop AI systems that show empathy [6] [7].

Despite encouraging results, there are three main challenges with the current studies. First, since publicly available datasets are limited, the models generalize to various demographic groups and cultural expressions. Second, ensemble techniques or combinations of models often have ununderutilized potential to improve performance. Lack of interpretability in

model predictions is another limitation; they can impede trust and adoption, especially in sensitive applications (e.g., healthcare or education). To overcome these shortcomings, we developed a custom 2D CNN model as well as utilized pretrained models such as VGG16, ResNet50, EfficientNetB0, Xception, DenseNet121, and MobileNetV2. In addition, we used ensemble learning, combining the best individual models such as Xception and MobileNetV2 to obtain better accuracy and robustness [8] [9].

In the domain of facial emotion classification, this study provides enriched literature on human emotion recognition through facial expressions. This indicates that transfer learning methods can improve the accuracy of deep learning models, especially when combined with ensemble methods for classifying emotions like happiness, sadness, anger, and surprise. By merging two datasets, the research builds a more generalized model, more capable of distinguishing subtle differences in emotions. Ensemble techniques increase accuracy by aggregating different models to overcome misclassification. These developments have resulted in more accurate tools for emotion recognition in cases such as psychology and human-computer interaction.

II. Related Works

Because of their different applications in healthcare, security, human-computer interaction, etc., Facial Emotion Recognition (FER) has been a key research domain. One of the most commonly used benchmarks for assessment of FER models is the extended Cohn-Kanade (CK+) dataset. In this section, major studies related to FER and related to CK+ are reviewed.

Noor et al. [10] developed a framework that utilizes handcrafted feature extraction techniques like SURF, FAST, HOG, and Harris Corner Detection. Feature quantization using supervised k-means clustering was applied to the preprocessed images using the Viola-Jones algorithm. They developed a model that correctly classified seven basic emotions with an accuracy of 90.79% on the CK+ dataset.

Chen et al. [11] introduced a hybrid method that uses geometry and different appearances based on an analysis between neutral faces and expressive faces. They proposed the use of SVM for classification and achieved 95% accuracy on CK+, which signals a the potentialse of integrating multi-features.

Ezerceli and Eskil [12] proposed a CNN-based FER system trained on merged datasets, including CK+. With an accuracy of 93.7%, their model demonstrates that training on a hybrid dataset that consists of generics and skin disease images increases the robustness of the model and tackles issues like intra-class variability.

Wafi et al. [13] compared features extractor methods (LBPs + facial landmarks) on the CK+ dataset. An adaptive extreme learning machine (aELM) with 88.07% accuracy provides a link between standard ELM methods and incremental processes. For FER, Borgalli and Surve [14] developed a custom CNN architecture using the FER13, CK+, and JAFFE datasets. The model reached 91.58% of correct detection for the seven basic emotions. K-fold cross validation was used in the study, which comes to show that custom architectures can be effective for emotion recognition.

Subramanian et al. For example, [15] investigated multimodal emotion recognition from a combination of facial, audio, and EEG data. Their model reached 71.24% accuracy on CK+ when applying fusion techniques, providing further proof to the advantage of using multiple different modalities for emotion recognition.

III. Materials and methods

A. Dala Collection and Analysis

The Extended Cohn-Kanade (CK+) dataset is one of the most widely used benchmarks for facial expression and emotion recognition research. The dataset contains a total of 920 sample grayscale images, each of size 48×48 pixels, representing the facial expression with a variety of emotions. The images are categorized into seven classes: anger, disgust, fear, happy, sadness, surprise, and contempt. The number of images in each category is listed in Table I. Fig. 2 illustrates some examples from the data.

TABLE I
NUMBER OF IMAGES PER CATEGORY

Category	Number of Images
Anger	135
Disgust	177
Fear	75
Happiness	207
Sadness	84
Surprise	249
Contempt	177
Total Images	**1104**

B. Train-Test Split

In this work, 80% images from the aforementioned dataset are randomly sampled for being used in the training process, and the rest 20% are used for model testing and evaluation. 10% images from the training set are further used for validation during the training process.

C. Feature Extraction

Algorithm 1 Feature Extraction

Require: Dataset D, Pretrained CNN model (InceptionV3, MobileNet)

Ensure: Extracted Features

1: D_{train} , D_{test} ← Split80to20 (D)
2: $F_{\text{train(InceptionV3)}}$ ← InceptionV3 (D_{train})
3: $F_{\text{train(EfficientNetB3)}}$ ← MobileNet (D_{train})
4: $F_{\text{train(Fused)}}$ ←Concatenate($F_{\text{train(InceptionV3)}}$, $F_{\text{train(MobileNet)}}$)
5: RETURN $F_{\text{train(InceptionV3)}}$, $F_{\text{train(MobileNet)}}$, $F_{\text{train(Fused)}}$

In this work, pretrained versions of two SOTA CNN models, namely InceptionV3 and MobileNet, are used. Here pretrained refers to the fact that the weights and biases of these models were achieved while training with Imagenet are

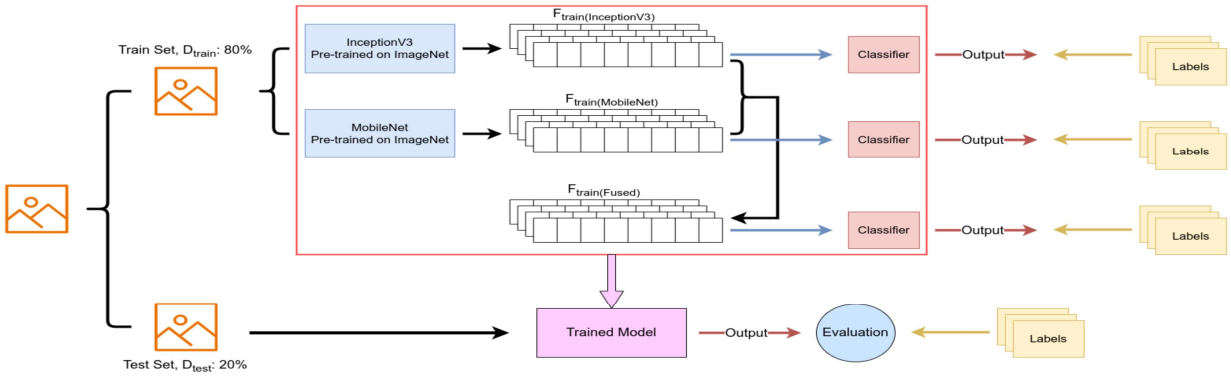

Fig. 1. Proposed Workflow

Fig. 2. Example Images from Dataset [16]

retained and used on our specific FER dataset. This approach works well enough to extract meaningful features from our dataset, which is new to the models. This approach works due to the fact that while being trained with a large image dataset like ImageNet, the models learn to extract important features from images in a very general way. Thus, we got two sets of features from InceptionV3 and MobileNet. We also fused these two sets of features to derive a new set of features in the hope that two different models will interpret the same image from slightly different perspectives, resulting in even more discriminating features. This approach is summarized in Algorithm 1.

D. Classification

Algorithm 2 Classification Workflow

Require: Feature-Set, Classifiers (Logistic Regression LG, Custom CNN $CCNN$, Voting Classifier V), Labels L
Ensure: Trained Classifiers $LG_{\text{trained}}, CCNN_{\text{trained}}, V_{\text{trained}}$

```
for each feature F in Feature-Set do
    for each classifier C in Classifiers do
        C.fit(F, L)
    end for
end for
return LG_trained, CCNN_trained, V_trained
```

Three sets of features, namely $F_{\text{train(InceptionV3)}}$, $F_{\text{train(MobileNet)}}$, $F_{\text{train(Fused)}}$ given as input to the classifier algorithms to train them. As classifiers, we have incorporated three approaches—

- **Logistic Regression:** A classifier based on probabilistic modeling that generates a probability distribution across the classes.
- **Custom CNN Model:** A tailored Convolutional Neural Network (CNN) consisting of three Conv1D layers with progressively increasing filter sizes (32, 64, 128), each followed by a MaxPooling1D layer to reduce dimensionality. The resulting output is flattened and passed through a dense layer with 256 neurons and a dropout rate of 0.5 for regularization, concluding with a softmax layer for final classification.
- **Voting Classifier:** An ensemble model that integrates predictions from multiple classifiers, including Logistic Regression, Linear Support Vector Machine (LSVM), and Perceptron. Additionally, it combines predictions from a broader set of classifiers, such as Logistic Regression, LSVM, Perceptron, SVM with quadratic penalties, and SVM with a smooth quadratic loss, utilizing a hard voting mechanism.

Among the experiments we have conducted, features extracted using MobileNet when classified with the voting classifier give the best performance with fewer parameters. This approach yields 98.48% classification accuracy with 98.57% of precision, 98.48% of recall, and 98.42% of f1-score. The overall classification procedure is depicted in Algorithm 2. The whole workflow is illustrated in Fig. 1.

IV. Result and Analysis

In this work, we have utilized two CNN models, namely InceptionV3 and MobileNet, with their pre-training on ImageNet to extract features. Extracted futures are then fused to derive yet another set of features. Then, all these sets of features are classified using three instances of classifiers, namely Logistic Regression, a custom 1D CNN, and the voting classifier. Among multiple experiments we have conducted, the voting classifier outperforms other experiments in terms of classification accuracy and other evaluation metrics, in spite of having fewer model parameters. Results of our experiments are summarized in Table II. The confusion matrix of our best approach is illustrated in Fig. 3.

V. Limitations and Future Scope

In spite of encouraging performance, our approach is plagued with some limitations. Firstly, the dataset we have

TABLE II
PERFORMANCE METRICS FOR DIFFERENT FEATURE EXTRACTORS AND CLASSIFIERS

Feature Extractor	Classifier	Acc	Macro Pre	Macro Rec	Macro F1	Weighted Pre	Weighted Rec	Weighted F1
InceptionV3, MobileNet	Logistic Regression	97.47	98.43	95.46	96.71	97.64	97.40	97.87
InceptionV3, MobileNet	Custom Classifier	98.48	99.19	97.14	98.00	98.57	98.48	98.42
InceptionV3, MobileNet	Voting Classifier	97.98	98.94	96.61	97.59	98.13	97.98	97.93
MobileNet	Logistic Regression	96.97	98.32	94.93	96.37	97.21	96.97	96.91
MobileNet	**Voting Classifier**	**98.48**	**99.19**	**97.14**	**98.00**	**98.57**	**98.48**	**98.42**

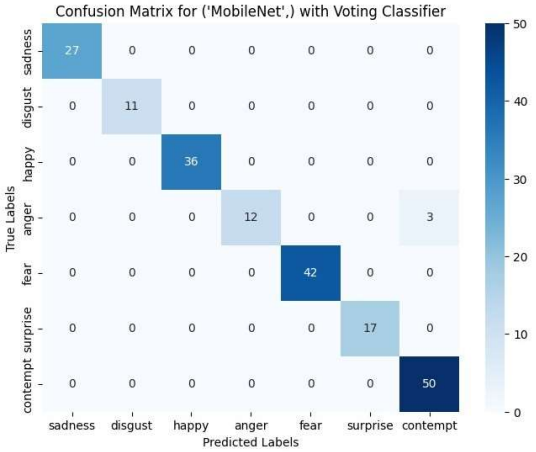

Fig. 3. Confusion Matrix

used contains low-resolution images (i.e., only 48x48 pixels). Moreover, the images are grayscale, which lacks important color information. Furthermore, the hyperparameters of the classifiers are not tuned to their optimal values. Besides, we have used pretrained CNNs as stand-still feature extractors that are not fine-tuned for our specific datasets.

In the future scope of this work, different datasets with higher resolution images and color information may be used to train the classifiers. Moreover, hyperparameters may be tuned to optimal values to increase classification accuracy. Feature-extracting CNNs may be tuned to our specific purpose to get more general and accurate models.

VI. CONCLUSION

In this study, we proposed a transfer learning-based approach for classifying FER images into six categories— anger, disgust, fear, happy, sadness, surprise, and contempt. We utilized two CNN models to extract features from our dataset by leveraging their pre-trained knowledge on the ImageNet dataset. The extracted features were then fused to create a combined feature set along with the original feature sets. These feature sets were classified using various classifiers, including Logistic Regression, a custom 1D CNN, and a hard voting classifier that integrates predictions from Logistic Regression, Linear Support Vector Machine (LSVM), perceptron, a quadratically penalized SVM and an SVM with a quadratically smooth loss. The feature set extracted using MobileNet, when classified using the voting classifier, achieved the highest performance with classification accuracy, precision, recall, and f1-score of 98.48%, 99.19%, 97.14%, 98.00%, 98.57%, 98.48%, and 98.42%, respectively. Further fine-tuning of the CNN models and hyperparameter optimization could potentially enhance classification accuracy. Unlike our dataset, a dataset with higher resolution and color information may increase the generality of the model. In summary, this work establishes a foundation for the accurate classification of FER images, aiding in classification of various human emotions from images.

REFERENCES

[1] P. Ekman and W. V. Friesen, "Constants across cultures in the face and emotion," *J. Pers. Soc. Psychol.*, vol. 17, no. 2, pp. 124–129, 1971.

[2] G. Zhao and M. Pietikäinen, "Dynamic texture recognition using local binary patterns with an application to facial expressions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 6, pp. 915–928, 2007.

[3] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "AffectNet: A database for facial expression, valence, and arousal computing in the wild," *IEEE Trans. Affect. Comput.*, vol. 10, no. 1, pp. 18–31, 2019.

[4] S. Li and W. Deng, "Deep Facial Expression Recognition: A Survey," *IEEE Trans. Affect. Comput.*, vol. 13, no. 3, pp. 1195–1215, 2022.

[5] B. C. Ko, "A brief review of facial emotion recognition based on visual information," *Sensors (Basel)*, vol. 18, no. 2, 2018.

[6] D. Kollias and S. Zafeiriou, "Exploiting deep learning for facial expression recognition," *Image and Vision Computing*, vol. 79, pp. 38–48, 2019.

[7] I. Kotsia and I. Pitas, "Facial expression recognition in image sequences using geometric deformation features and Support Vector Machines," *IEEE Trans. Image Process.*, vol. 16, no. 1, pp. 172–187, 2007.

[8] J. Kim and H. Kang, "Ensemble Learning for Emotion Recognition from Facial Expressions," *Expert Systems with Applications*, vol. 184, 2021.

[9] H. Guo and X. Sun, "Improving Facial Emotion Recognition with Transfer Learning and Explainability Techniques," *Applied Intelligence*, vol. 52, no. 3, pp. 1234–1245, 2022.

[10] J. Noor, M. Daud, R. Rashid, H. Mir, S. Nazir, and S. A. Velastin, "Facial expression recognition using hand-crafted features and supervised feature encoding," in *2020 International Conference on Electrical, Communication, and Computer Engineering (ICECCE)*, 2020.

[11] J. Chen, D. Chen, Y. Gong, M. Yu, K. Zhang, and L. Wang, "Facial expression recognition using geometric and appearance features," in *Proceedings of the 4th International Conference on Internet Multimedia Computing and Service*, 2012.

[12] O. Ezerceli and M. T. Eskil, "Convolutional neural network (CNN) algorithm based facial emotion recognition (FER) system for FER-2013 dataset," in *2022 International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME)*, 2022.

[13] M. Wafi, F. A. Bachtiar, and F. Utaminingrum, "Feature extraction comparison for facial expression recognition using adaptive extreme learning machine," *International Journal of Electrical & Computer Engineering*, 2023.

[14] R. A. Borgalli and S. Surve, "Custom CNN Architecture for FER Using FER13, CK+, and JAFFE Datasets," *Journal of Physics: Conference Series*, vol. 2236, no. 1, 2022.

[15] G. Subramanian, N. Cholendiran, K. Prathyusha, N. Balasubramanain, and J. Aravinth, "Multimodal emotion recognition using different fusion techniques," in *2021 Seventh International Conference on Bio Signals, Images, and Instrumentation (ICBSII)*, 2021.

[16] Kaggle.com. [Online]. Available: *https://www.kaggle.com /datasets/shawon10/ckplus* [Accessed: 15-Jan-2025].