Facial Emotion Recognition using OpenCV and Transfer Learning

Md. Intikhab Shahriar Hasan, MSc Student, *FH-2088*

Department of Electrical and Electronic Engineering

University of Dhaka, Dhaka-1000, Bangladesh

***Abstract*—Facial Emotion Recognition (FER) has emerged as a pivotal research domain in human-computer interaction, enabling advancements in fields like psychology, healthcare, and artificial intelligence. This study leverages transfer learning using the VGG19 architecture for emotion classification on four benchmark datasets: JAFFE, CK+, KDEF, and FER-2013, achieving accuracies of 97.67%, 100%, 95.07%, and 74.45%, respectively. The methodology integrates OpenCV-based preprocessing, involving image cropping, resizing, normalization, and augmentation through random rotations and translations. The modified VGG19 network incorporates a customized classifier with frozen convolutional layers to extract salient features effectively. The model was trained using the Adam optimizer with a learning rate of 0.0001, over 20 epochs, and with a batch size of 8. Results were validated using an 80-20 train-test split. Comparative analysis highlights the model's competitive performance, surpassing or equaling state-of-the-art accuracies reported in recent works. For instance, the study outperforms references employing Convolutional Neural Networks (CNN) enhanced with attention mechanisms, ELM classifier, Vision Transformers, and saliency maps integrated with advanced preprocessing techniques like GANs and CLAHE. These findings underline the efficacy of the proposed methodology and its potential contribution to the development of robust FER systems.**

Keywords—Facial Emotion Recognition (FER), Human-Computer Interaction (HCI), OpenCV, Convolutional Neural Networks (CNNs), Fully Connected Neural Networks (FCNNs), Long Short-term Memory (LSTM), Transfer Learning (TL), VGG19, JAFFE, CK+, KDEF, FER-2013.

I. INTRODUCTION

Facial Emotion Recognition (FER) represents a multidisciplinary frontier, integrating advancements in computer vision, artificial intelligence, and psychology to interpret human emotions from facial expressions. The ability of machines to recognize and classify emotions has transformative implications across domains such as human-computer interaction, healthcare, security, marketing, and education. Despite substantial progress, FER remains a challenging task, primarily due to inherent variations in facial expressions caused by differences in lighting conditions, head poses, occlusions, and individual-specific traits. [1]. Human facial expressions are a rich source of information, conveying emotional states that form the backbone of non-verbal communication. According to Ekman's theory of basic emotions, universally recognized emotions, namely happiness, sadness, anger, fear, surprise, disgust, and neutral, manifest through consistent facial expressions, shown in Fig. 1. Consequently, automating FER

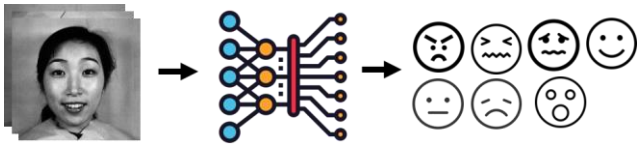


Fig. 1. Emotion recognition from human facial images with the help of machine learning techniques.

has gained traction for applications in affective computing, mental health diagnostics, user experience design, and real-time surveillance systems. The burgeoning interest in FER is also evident in its market potential, with the global emotion recognition market projected to grow significantly, reflecting its rising demand and expanding applicability [2]. While FER has evolved significantly with the advent of deep learning, it faces notable obstacles [1], [3]:

- *Data Variability:* Differences in image resolution, illumination, head orientation, and facial occlusions add complexity to emotion classification.
- *Class Imbalance:* Limited samples for certain emotions within datasets hinder the development of generalizable models.
- *Real-World Scenarios:* Models trained in controlled environments often struggle with real-world data due to noise and unpredictability.
- *Efficiency vs. Accuracy Trade-off:* Achieving high accuracy while maintaining computational efficiency is critical, particularly for real-time applications.

The transition from traditional machine learning methods to deep learning architectures has revolutionized FER. Convolutional Neural Networks (CNNs), Vision Transformers (ViTs), and hybrid models have demonstrated superior performance by automating feature extraction and learning hierarchical representations of facial features [3]. Recent research has introduced novel preprocessing and data

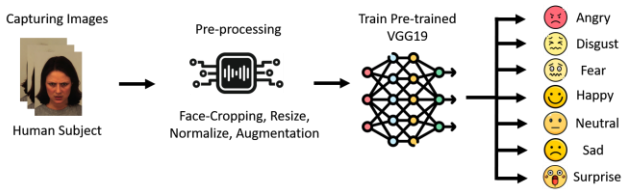


Fig. 2. Block diagram of the proposed system. The system preprocesses images to crop the face, resize, normalize and augment them. Following, they are fed to pre trained VGG19 network to fine-tune it for the Facial Emotion Recognition (FER) task.

augmentation techniques, such as saliency mapping, generative adversarial networks (GANs), and contrast-limited adaptive histogram equalization (CLAHE), to mitigate challenges like poor image quality and class imbalance [4], [2]. Transfer learning, which leverages pre-trained networks, has emerged as a potent approach to enhance performance while reducing computational overhead [1], [2], [3]. This research employs transfer learning using the VGG19 architecture, a deep CNN known for its robustness and feature extraction capabilities [5], to classify emotions on four widely used FER datasets: JAFFE, CK+, KDEF, and FER-2013 [2], [3]. Fig. 2 presents the block diagram of our proposed approach. Preprocessing steps—including image resizing, normalization, and augmentation—are meticulously executed using OpenCV [6], [7] to ensure consistency and robustness. By freezing the feature extraction layers of VGG19 and fine-tuning the classifier, the model achieves state-of-the-art accuracies, outperforming several contemporary methodologies.

The contributions of this work are as follows:

1. *Comprehensive Preprocessing Pipeline:* The use of OpenCV for image resizing, cropping, normalization, and augmentation ensures high-quality input data. Augmentation techniques, such as random rotations and translations, enhance dataset diversity and mitigate overfitting.
2. *Optimized Training Strategy:* Employing the Adam optimizer with a learning rate of 0.0001, using dropout of 0.2 before final dense layers, and incorporating model checkpoints guarantees efficient learning while retaining the best model weights.
3. *Robust Performance Across Benchmarks:* The proposed approach demonstrates superior or competitive accuracies compared to recent studies on JAFFE, CK+, KDEF, and FER-2013 datasets.
4. *Detailed Comparative Analysis:* Results are contextualized against recent state-of-the-art methods, including CNNs enhanced with attention mechanisms, Vision Transformers, and GAN-based preprocessing frameworks.

The remainder of this paper is organized as follows: Section II reviews relevant literature, highlighting contemporary challenges and advancements in FER. Section III details the methodology, including preprocessing, network architecture modification, and training methods. Section IV presents experimental results and performance comparisons. Section V discusses implications, limitations, and potential future directions, and finally concludes the study by summarizing key findings and contributions.

By addressing key challenges and leveraging advanced techniques, this study aims to contribute to the growing body of knowledge in FER and inspire further innovations in this dynamic field.

II. Related Works

Facial Emotion Recognition (FER) has undergone substantial advancements, transitioning from handcrafted feature-based methods to data-driven deep learning approaches. This section reviews significant contributions to FER, categorized into traditional methods, deep learning techniques, hybrid approaches, and dataset-specific advancements.

A. Traditional Methods for FER

Early FER systems relied on handcrafted features and conventional machine learning classifiers. Techniques such as Local Binary Patterns (LBP), Histogram of Oriented Gradients (HOG), and Scale-Invariant Feature Transform (SIFT) were widely employed for feature extraction, while Support Vector Machines (SVM), Random Forest (RF), and k-Nearest Neighbors (kNN) served as classifiers. Although computationally efficient, these methods were limited by their inability to generalize across diverse datasets, as they depended heavily on manual feature engineering and were sensitive to variations in lighting, pose, and occlusion [2], [3].

For instance, early work by J. Wang and Y. Hong-mei employed the JAFFE dataset and demonstrated the potential of LBP for emotion classification, achieving reasonable accuracies under controlled settings [8]. Similarly, N. Zeng, H. Zhang, B. Song, W. Liu, Y. Li, and A. M. Dobaie explored the use of HOG features combined with a shallow CNN to classify facial expressions in the CK+ dataset, highlighting the limitations of traditional methods when faced with real-world variability [9].

B. Emergence of Deep learning in FER

The advent of deep learning marked a paradigm shift in FER, enabling automated feature extraction and hierarchical representation learning. Convolutional Neural Networks (CNNs) emerged as the primary architecture for FER tasks, offering significant improvements in accuracy and robustness. Studies leveraging pre-trained networks, such as AlexNet, VGGNet, and ResNet, demonstrated the efficacy of transfer learning in reducing the need for extensive labeled datasets [10].

Khattak et al. proposed a deep CNN-based FER model, achieving accuracies of 95.65% on the JAFFE dataset and 99.36% on CK+ [11]. Their work underscored the importance of optimizing network architecture and hyperparameters to improve performance. Similarly, Kumari et al. introduced a saliency map and GAN-based preprocessing framework [4], coupled with a Nadam-optimized CNN, achieving state-of-the-art results on the JAFFE, CK+, and FER-2013 dataset. These

studies highlight the critical role of data preprocessing and model optimization in enhancing FER performance.

C. Hybrid Approaches and Attention Mechanisms

Recent research has focused on integrating attention mechanisms and hybrid models to address the challenges of class imbalance and feature discrimination. Vision Transformers (ViTs) and hybrid CNN-ViT architectures have been employed to capture both local and global features, significantly improving classification accuracy. For example, a study by Q. Huang et al. introduced a grid-wise attention mechanism employed with a Visual Transfomer [12], enhancing both low-level feature extraction and high-level semantic representation, achieving remarkable accuracy on the CK+ and FER+ datasets.

In parallel, ARBEx [13], a framework driven by Vision Transformers and reliability balancing, addressed class distribution issues in the FER task using JAFFE, FER+ and RAF-DB datasets. The incorporation of attention mechanisms allowed for dynamic weighting of features, enabling the model to focus on the most salient regions of the face.

D. Dataset Specific Innovations

FER datasets vary widely in their characteristics, from controlled lab environments (e.g., CK+, JAFFE) to real-world scenarios (e.g., FER-2013, FER+, RAF-DB, AffectNet) [2], [3]. This variability has driven dataset-specific innovations in preprocessing, augmentation, and model design. For instance, in another study of Kumari et al. employed CLAHE with modified joint trilateral filter is applied to the obtained enhanced images from FER+, JAFFE, and CK+ datasets to remove the impact of impulsive noise [14] before feeding them to deep neural network. Similarly, N. Banskota et al. used CNNs for feature extraction from images and Extreme Learning Machine (ELM) to classify emotion categories on CK+ and JAFFE datasets [15].

The progression from traditional methods to deep learning and hybrid approaches has significantly advanced FER. However, challenges such as real-world applicability, efficiency, and fairness remain. This study builds upon these advancements by leveraging VGG19-based transfer learning and image preprocessing to achieve state-of-the-art performance, setting a strong foundation for further research in FER.

III. Methodology

The methodology adopted in this study for Facial Emotion Recognition (FER) revolves around transfer learning using the VGG19 architecture, a widely recognized Convolutional Neural Network (CNN) model known for its feature extraction prowess. This section delineates the systematic approach, encompassing preprocessing, network architecture modifications, training procedures, and evaluation metrics, to ensure a robust and reproducible framework.

A. Dataset Description

This study leverages the utilization of four benchmark datasets for facial emotion recognition. The JAFFE dataset was introduced in 1996 by Miyuki G. Kamachi, Michael J. Lyons, and Jiro Gyoba at the Kyushu University in Japan as part of a study in emotion recognition. It contains 213 facial expression images of 10 Japanese female subjects, each expressing 7 distinct emotional states. [16, 17].

The CK+ dataset, developed by Peter J. Cohn and colleagues at the University of Pittsburgh, is an extension of the original Cohn-Kanade dataset (CK) and includes a broader set of facial expressions, as well as a larger and more diverse set of participants. It was introduced in 2003 and has since been widely adopted for emotion recognition research. The extended version, CK+, was released in 2010, offering improved data and labeling. It consists of 1237 images captured from 123 subjects (74 female and 49 male) [18].

The KDEF dataset was created by the Karolinska Institute (specifically by researchers including Lars L. Lundqvist, Anders

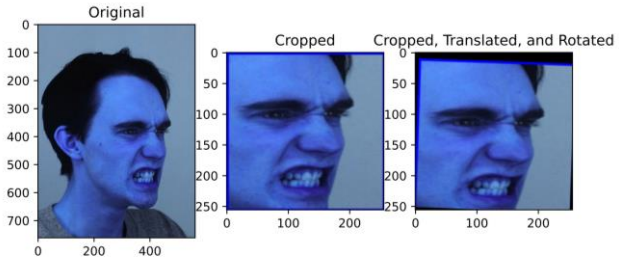


Fig. 3. The preprocessing of images, which includes cropping the face, resizing, normalizing, and augmenting them. Following, they are fed to the VGG19 network. This example uses a sample image from the KDEF dataset.

Flykt, and Mikael Öhman) to provide a set of facial expression images that could be used to investigate emotion recognition in a controlled manner. The dataset was first introduced in the early 2000s and has since become one of the most prominent datasets for emotion recognition tasks. KDEF contains images of 70 different individuals (35 male and 35 female) displaying facial expressions corresponding to 7 distinct emotions, including neutral expressions, totaling 4900 images. Each of the 7 emotions is presented under different viewing angles (frontal, left, right, and 45-degree rotated views). The dataset is designed to facilitate studies on both static facial expression recognition (i.e., emotion classification from a single image) and dynamic emotion recognition (i.e., understanding how emotions evolve in sequences) [19].

The FER-2013 dataset is a large-scale dataset designed for the task of facial expression recognition. It was released as part of the ICML 2013 Challenge and consists of a collection of facial expression images annotated with emotion labels. The dataset was created by researchers at The University of Toronto and is publicly available for research purposes. The FER-2013 dataset consists of a total of 35,887 images, each with a 48x48 pixel resolution (grayscale), which is a key aspect of its design. The small image size reduces computational complexity and memory requirements, making it practical for training machine learning models. The dataset includes 7 distinct emotion categories based on the basic emotions model. The dataset contains both male and female faces from various ethnic groups, although it is predominantly composed of images from

Caucasian and Asian individuals. This limited cultural diversity can be a consideration when applying models in cross-cultural settings. The subjects in the dataset come from a range of age groups, including children, adults, and elderly individuals, allowing for some variation in age-related facial features and expression characteristics [20].

All the above datasets have been utilized in a range of studies related to affective computing, human-computer interaction, and emotion recognition. They are pivotal in our study of building a robust facial emotion recognition system.

B. Data Preprocessing

Preprocessing is critical in ensuring data consistency, enhancing feature extraction, and mitigating overfitting. The study leverages the OpenCV library for image handling and augmentation, with steps detailed below:

Cropping and Resizing: Facial regions were extracted using Haar cascade classifiers to ensure the focus remains on facial features [21]. The cropped facial regions were resized to 256x256 pixels to align with the input requirements of the VGG19 model. Pixel values were normalized to the [0, 1] range by dividing by 255. This step ensures uniformity across datasets and accelerates convergence during training.

Data Augmentation: Augmentation techniques such as random rotations (up to 15 degrees), and translations (up to 15 pixels) were employed [22]. Augmentation was applied exclusively to the training set to improve model generalizability and reduce overfitting. The outcome of face-cropping and augmentation can be found in Fig. 3.

C. Network Architecture & Training

The Transfer Learning Model: The VGG19 model, pre-trained on the ImageNet dataset [5], was adapted for FER task. All convolutional layers were frozen to retain the pre-trained feature extraction capabilities of VGG19. The fully connected layers of VGG19 were replaced with a new classification head tailored to FER. After the base model, Dropout (rate: 0.2) for regularization was added before flattening, followed by two dense layers with 256 and 7 (number of classes) neurons, respectively, each followed by ReLU activation. A final dense layer with Softmax activation, outputting probabilities for the emotion classes.

Training Parameters: The Adam optimizer with a learning rate of 0.0001 was employed for training. Categorical Cross-Entropy was used as the loss function to handle multi-class classification. An 80-20 split was used to partition the datasets into training and testing subsets. Stratified sampling ensured class distribution parity across the subsets. A batch size of 8 was chosen to balance memory constraints and training efficiency. Training was conducted over 20 epochs, the best model weights were saved based on validation accuracy, ensuring optimal performance. The final model was evaluated on the test set using metrics such as accuracy, precision, recall, F1-score, and confusion matrices to provide a comprehensive assessment.

Experimental Setup: Experiments were conducted on Kaggle's data science platform with an NVIDIA's P100 GPU, and 29GB of RAM. TensorFlow and Keras were used for model implementation. The OpenCV library facilitated image preprocessing and augmentation.

D. Performance Evaluation

Performance evaluation shows the effectiveness and reliability of machine learning models. In this study, the evaluation was conducted using standard metrics and visualization techniques to assess the classification performance. This section provides a comprehensive explanation of the metrics and methods used to evaluate the models.

Accuracy: Accuracy measures the proportion of correctly classified samples among the total samples. It is the primary metric used to compare model performance in this study:

$$Accuracy = \frac{Correct\ Predictions}{Total\ Predictions} \quad (1)$$

While accuracy is a useful metric, it can be misleading for imbalanced datasets, where certain classes may dominate. In our case, the TESS dataset has a balanced distribution across all emotions, making accuracy an appropriate measure.

Precision, Recall, and F1 Score: To further evaluate model performance for each class, precision, recall, and F1-score were calculated. Precision is the ratio of correctly predicted positive observations to the total predicted positive observations. Recall or Sensitivity is the ratio of correctly predicted positive observations to all actual positive observations. F1-Score is the harmonic mean of precision and recall, providing a single measure of a model's performance. They are provided as:

$$Precision = \frac{True\ Positives\ (TP)}{True\ Positives(TP) + False\ Positives\ (FP)}$$

$$Recall = \frac{True\ Positives\ (TP)}{True\ Positives(TP) + False\ Negatives\ (FN)}$$

$$F1\ Score = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (2)$$

Confusion Matrix: A confusion matrix provides a detailed breakdown of model predictions for each class, showing:

- True Positives (TP): Correctly predicted samples of a class.
- False Positives (FP): Samples incorrectly predicted as a class.
- False Negatives (FN): Samples of a class incorrectly predicted as another class.
- True Negatives (TN): Samples correctly excluded from a class.

TABLE I
MODEL'S PERFORMANCE ON DIFFERENT DATASETS

Dataset	Accuracy (%)	Precision	Recall	F1-Score
JAFFE	97.67	97	98	97
CK+	100	100	100	100
KDEF	95.07	95	95	95
FER-2013	74.45	78	73	75

The confusion matrix highlights areas where the model struggles, enabling targeted improvements [23].

IV. Results & Discussions

This section presents the outcomes of the experiments conducted to evaluate the performance of the proposed facial emotion recognition system using the JAFFE, CK+, KDEF, and FER-2013 datasets. The results are discussed in terms of classification accuracy, and a confusion matrix is provided for the best-performing model. The training and testing accuracy and loss vs epochs trained are shown to demonstrate the learning behavior of the neural network architecture. This section also provides a detailed comparison to prior studies at the end.

A. *Classification Results*

To train the VGG19 on different datasets for facial emotion recognition, we provided the resized and face-cropped images, augmented with random rotations and translation. The 80-20 stratified train-test splitting with random state = 42 was carried out and the labels vector was one-hot encoded. The following Table I shows the performance of the VGG19 in the given task based on classification accuracy.

The model achieved an accuracy of 97.67% on JAFFE dataset, which is 1% ahead of [15]'s performance on the same dataset. On the CK+ dataset, it outperformed all other prior works, acquiring a 100% score. This result demonstrates the efficacy of pre-trained deep-learning image classification networks in extracting and learning hierarchical features from

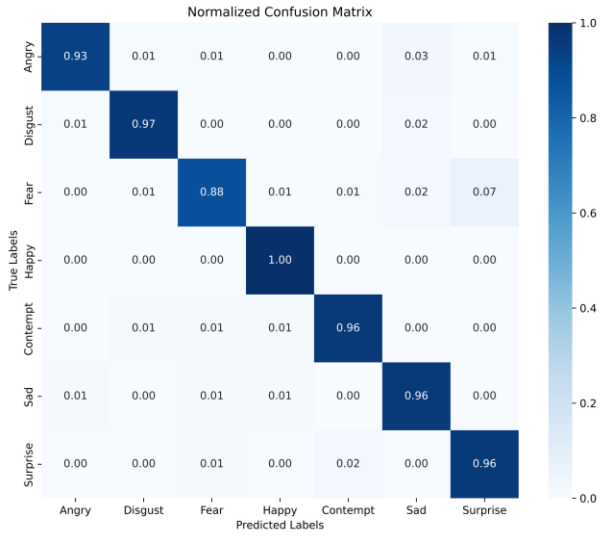
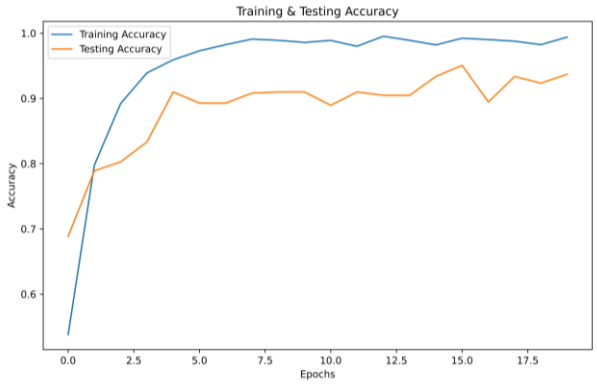


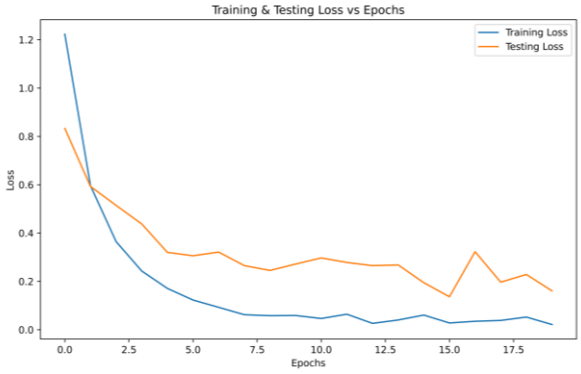
Fig. 4. The confusion matrix of VGG19 on KDEF dataset, classifying 7 different emotion categories from human facial images. The dataset was split into 80-20 train-test split and the classification accuracy was 95.07%.

facial data. Following the evaluation of these two benchmark datasets, we focused on two other benchmark datasets, KDEF and Fer-2013.

The prior study of Erlangga et al [24]. shows 93% accuracy on the KDEF dataset using transfer learning with Inception v3. We were able to get a 95.07% accuracy score on the same dataset using our model. Lastly, we evaluated our network's performance on a subset of the FER-2013 dataset due to



(a)



(b)

Fig. 5. The training and validation (a) Accuracy vs epochs trained and (b) Loss vs epochs for VGG19, trained on the KDEF dataset. curves demonstrated convergence, with no signs of overfitting or underfitting.

resource constraints, where we kept just over 1500 image samples for each of the emotion classes, and trained our model using the same methodology discussed above. It showed an accuracy of 74.45%, which couldn't surpass [4]'s score of 84.2%. With more resources available, our proposed model's performance is subjected to be evaluated on the full set of FER-2013 dataset. The Fig. 4 shows confusion matrix of VGG19's performance on the KDEF dataset. The training and validation loss curves in Fig. 5 demonstrated smooth convergence, with no signs of overfitting or underfitting. This consistency indicates that the model is not biased toward any specific class, a common issue in many classification tasks.

B. *Comparison with Prior Studies*

The results of this study were compared to six recent works that achieved state-of-the-art performance using various deep learning techniques on these four benchmark datasets [7], given in Table II. On top of that, our work demonstrates the following advancements:

1. *Improved Accuracy:* The proposed methodology achieved state-of-the-art accuracy on the JAFFE

(97.67%) and CK+ (100%) datasets, demonstrating its robustness in controlled environments with well-labeled data. Performance on KDEF (95.07%) and FER-2013 (74.52%) highlighted the model's ability to generalize to datasets with increased complexity, including varying angles, lighting, and background noise. The relatively lower accuracy on FER-2013 can be attributed to its real-world nature, where data exhibits significant variability in quality and emotional intensity.

2. *Simplified Preprocessing:* The preprocessing pipeline, including cropping, resizing, normalization, and augmentation, was pivotal in enhancing model performance. Augmentation techniques, such as random rotations and translations, enriched the diversity of training data, mitigating overfitting and improving generalization. Normalization ensured consistent input ranges, accelerating convergence and stabilizing training dynamics.
3. *Efficiency:* The design of our transfer learning method has less processing chain, optimized for avoiding computational burden, making it suitable for real-time emotion recognition applications. Freezing the convolutional layers of VGG19 preserved its feature extraction capabilities, and faster learning, while the custom classifier tailored the model for FER tasks. The integration of Dropout layers (rate: 0.2) in the classifier head effectively reduced overfitting, as evidenced by the alignment of training and validation accuracy curves.

C. Discussion

Unlike traditional models, the pre-trained CNNs effectively capture hierarchical relationships and temporal dependencies within the facial image features, enabling it to differentiate between emotions categories. Hyperparameter tuning, including the choice of freezing initial layers, options of regularization techniques, activation functions, and learning rate, played a crucial role in optimizing the model. Besides, our method relied solely on simple image preprocessing techniques such as face-cropping, resizing and normalizing, and augmentation with random rotation and translation. This highlights the efficiency of these steps when paired with a well-designed pre-trained model.

D. Ethical Considerations

Emotion recognition systems must be designed with privacy and ethical considerations in mind, particularly when deployed in sensitive applications like surveillance or mental health monitoring. Furthermore, to mitigate bias and fairness, ensuring that models perform equitably across diverse populations and environments is crucial to prevent potential biases in emotion detection.

V. Conclusion

This section delves into an in-depth discussion of the experimental findings, contextualizing the results in the broader scope of Facial Emotion Recognition (FER). The implications of the results, their alignment with existing literature, limitations, and potential avenues for further research are critically analyzed.

TABLE II
COMPARISON WITH RECENT RELEVANT WORKS

Reference & Year	Methodology	Dataset	Accuracy (%)
[11], 2021	CNN	CK+, JAFFE	99.36, 95.65
[4], 2023	Saliency Map+CLAHE +GAN+CNN	JAFFE, CK+, FER-2013	99.7, 99.9, 84.2
[15], 2022	CNN+ELM Classifier	JAFFE, CK+	96.67, 98.40
[24], 2024	Inception v3	JAFFE, KDEF	94, 93
This work	OpenCV image processing+ VGG19	JAFFE, CK+, KDEF, FER-2013	97.67, 100, 95.07, 74.45

A. Contributions of the Study

This study demonstrates the potential of transfer learning using VGG19 for FER, achieving state-of-the-art performance on benchmark datasets and providing valuable insights into the role of preprocessing and architecture modifications. While challenges persist in handling real-world data variability and computational efficiency, the proposed methodology offers a robust foundation for future advancements in FER. By addressing identified limitations and leveraging emerging technologies, this work paves the way for developing more accurate, fair, and efficient emotion recognition systems. Ultimately, these systems hold the promise of transforming human-computer interaction and enabling innovative applications across diverse domains.

B. Practical Applications

The findings of this study have significant implications for real-world applications, including [1], [2], [3]:

- Human-Computer Interaction: Emotion-aware systems can enhance user experience in virtual assistants, gaming, and educational tools.
- Mental Health Monitoring: Facial emotion recognition systems can be used to monitor emotional well-being, providing early detection of conditions like depression or anxiety.
- Customer Service: Emotion recognition can improve automated customer support by tailoring responses based on the customer's emotional state.

Its adaptability to controlled (e.g., CK+ and JAFFE) and real-world datasets (e.g., FER-2013) makes it suitable for diverse deployment scenarios. Furthermore, the model's modular architecture facilitates integration with hybrid approaches, such as Vision Transformers and attention mechanisms, to further enhance accuracy and robustness.

C. Challenges and Limitations

Our study includes the JAFFE and CK+ datasets, which consist of controlled, studio-quality samples with a balanced class distribution. Real-world datasets like KDEF and FER-2013, often include noise, speaker variability, and imbalanced classes, affects the system's performance. The FER-2013 dataset's high intra-class variability and low-resolution images posed significant challenges, limiting the model's performance. Angle variations in KDEF introduced inconsistencies in feature representation, impacting the precision of emotions like anger and fear. Furthermore, while JAFFE provided high-quality data, its lack of demographic diversity (e.g., gender, ethnicity) restricts the generalizability of findings to broader populations. Besides, though pre-trained CNNs are comparatively efficient, further optimization is required for deployment in low-resource environments, such as embedded systems. Despite its limitations, this research provides a foundation for advances in speech emotion recognition and sets the stage for addressing real-world challenges in future work.

D. Future Directions

The system's feasibility of implementation in real-world scenarios requires evaluating its robustness on datasets with real-world variability, such as crowded environments, cultural diversity, and spontaneous images. Extension of the study to include datasets in different racial and cultural contexts is preferred to ensure generalizability across diverse populations. Combining facial data with speech (voice recordings) and textual (written content) modalities can improve performance in ambiguous or overlapping emotion scenarios. Developing more lightweight architectures that retain high accuracy while minimizing computational complexity can leverage the model's deployability in edge and mobile devices. To better capture long-term temporal dependencies, especially if the data is sequences of images or video, incorporating recurrent neural networks (RNNs), Vision Transformers or attention mechanisms may help better capture long-term dependencies.

VI. References

[1] T. Kopalidis, Vassilios Solachidis, N. Vretos, and P. Daras, "Advances in Facial Expression Recognition: A Survey of Methods, Benchmarks, Models, and Datasets," *Information*, vol. 15, no. 3, pp. 135–135, Feb. 2024, doi: https://doi.org/10.3390/info15030135.

[2] C. Dalvi, M. Rathod, S. Patil, S. Gite, and K. Kotecha, "A Survey of AI-Based Facial Emotion Recognition: Features, ML & DL Techniques, Age-Wise Datasets and Future Directions," *IEEE Access*, vol. 9, pp. 165806–165840, 2021, doi: https://doi.org/10.1109/access.2021.3131733.

[3] A. R. Khan, "Facial Emotion Recognition Using Conventional Machine Learning and Deep Learning Methods: Current Achievements, Analysis and Remaining Challenges," *Information*, vol. 13, no. 6, p. 268, May 2022, doi: https://doi.org/10.3390/info13060268.

[4] N. Kumari and R. Bhatia, "Saliency map and deep learning based efficient facial emotion recognition technique for facial images," Jul. 2023, doi: https://doi.org/10.1007/s11042-023-16220-0.

[5] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *arXiv.org*, Apr. 10, 2015. https://arxiv.org/abs/1409.1556.

[6] R. T. Hasan and A. B. Sallow, "Face Detection and Recognition Using OpenCV," *Journal of Soft Computing and Data Mining*, vol. 2, no. 2, pp. 86–97, Oct. 2021, Available: https://publisher.uthm.edu.my/ojs/index.php/jscdm/article/view/8791.

[7] A. Sharma, J. Pathak, M. Prakash, and J. N. Singh, "Object Detection using OpenCV and Python," *IEEE Xplore*, Dec. 01, 2021. https://ieeexplore.ieee.org/document/9725638.

[8] J. Wang and Y. Hong-mei, "Face Detection Based on Template Matching and 2DPCA Algorithm," Jan. 2008, doi: https://doi.org/10.1109/cisp.2008.270.

[9] N. Zeng, H. Zhang, B. Song, W. Liu, Y. Li, and A. M. Dobaie, "Facial expression recognition via learning deep sparse autoencoders," *Neurocomputing*, vol. 273, pp. 643–649, Jan. 2018, doi: https://doi.org/10.1016/j.neucom.2017.08.043.

[10] M. A. H. Akhand, S. Roy, N. Siddique, M. A. S. Kamal, and T. Shimamura, "Facial Emotion Recognition Using Transfer Learning in the Deep CNN," *Electronics*, vol. 10, no. 9, p. 1036, Apr. 2021, doi: https://doi.org/10.3390/electronics10091036.

[11] A. Khattak, M. Z. Asghar, M. Ali, and U. Batool, "An efficient deep learning technique for facial emotion recognition," *Multimedia Tools and Applications*, Oct. 2021, doi: https://doi.org/10.1007/s11042-021-11298-w.

[12] Q. Huang, C. Huang, X. Wang, and F. Jiang, "Facial expression recognition with grid-wise attention and visual transformer," *Information Sciences*, vol. 580, pp. 35–54, Nov. 2021, doi: https://doi.org/10.1016/j.ins.2021.08.043.

[13] Wasi, Azmine Toushik and Rafi, Taki Hasan and Islam, Raima and Šerbetar, Karlo and Chae, Dong-Kyu, "ARBEx: Attentive Feature Extraction with Reliability Balancing for Robust Facial Expression Learning", *Computer Vision – ACCV 2024*, doi: https://doi.org/10.48550/arXiv.2305.01486.

[14] N. Kumari and R. Bhatia, "Efficient facial emotion recognition model using deep convolutional neural network and modified joint trilateral filter," *Soft Computing*, Feb. 2022, doi: https://doi.org/10.1007/s00500-022-06804-7.

[15] N. Banskota, A. Alsadoon, P. W. C. Prasad, A. Dawoud, T. A. Rashid, and O. H. Alsadoon, "A novel enhanced convolution neural network with extreme learning machine: facial emotional recognition in psychology practices," *Multimedia Tools and Applications*, Aug. 2022, doi: https://doi.org/10.1007/s11042-022-13567-8.

[16] M. J. Lyons, M. Kamachi, and J. Gyoba, "Coding Facial Expressions with Gabor Wavelets (IVC Special Issue)," *Zenodo*, Sep. 2020, doi: https://doi.org/10.5281/zenodo.4029680.

[17] 'Excavating AI' Re-excavated: Debunking a Fallacious Account of the JAFFE Dataset," *zenodo.org*, doi: https://doi.org/10.5281/zenodo.5147170.

[18] 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," *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, Jun. 2010, doi: https://doi.org/10.1109/cvprw.2010.5543262.

[19] Lundqvist, D., Flykt, A., & Öhman, A. (1998). The Karolinska Directed Emotional Faces - KDEF, CD ROM from Department of Clinical Neuroscience, Psychology section, Karolinska Institutet, ISBN 91-630-7164-9.

[20] I. J. Goodfellow *et al.*, "Challenges in representation learning: A report on three machine learning contests," *Neural Networks*, vol. 64, pp. 59–63, Apr. 2015, doi: https://doi.org/10.1016/j.neunet.2014.09.005.

[21] GeeksforGeeks, "Cropping Faces from Images using OpenCV Python," *GeeksforGeeks*, Jan. 13, 2021. https://www.geeksforgeeks.org/cropping-faces-from-images-using-opencv-python/

[22] "Image Transformations using OpenCV in Python," *GeeksforGeeks*, Dec. 01, 2022. https://www.geeksforgeeks.org/image-transformations-using-opencv-in-python/

[23] M. E. H. Chowdhury *et al.*, "Can AI Help in Screening Viral and COVID-19 Pneumonia?," *IEEE Access*, vol. 8, pp. 132665–132676, 2020, doi: https://doi.org/10.1109/access.2020.3010287

[24] E. S. Agung, A. P. Rifai, and T. Wijayanto, "Image-based facial emotion recognition using convolutional neural network on emognition

dataset," *Scientific Reports*, vol. 14, no. 1, p. 14429, Jun. 2024, doi: https://doi.org/10.1038/s41598-024-65276-x