Enhanced Emotion Recognition on the FER-2013 Dataset by Training VGG from Scratch

Muhammad Talha Jahangir[1], Syed Yameen Ali[1], Ali Hasnain[2], Muhammad Faizan[1], Mohsin Jabbar[3], Muhammad Kashan Basit[1], Abdul Rafay[1]
[1]Department of Computer Science, MNS-University of Engineering and Technology, Multan, Pakistan.
[2]College of Electrical and Mechanical Engineering, NUST, Pakistan
[3]Department of Artificial Intelligence NASTP Institute of Information Technology, Lahore, Pakistan.
*Correspondence: mtalhajahangir@mnsuet.edu.pk

Citation | Jahangir. M. T, Ali. S. Y, Hasnain. A, Faizan. M, Jabbar. M, Basit. M. K, Rafay. A, "Enhanced Emotion Recognition on the FER-2013 Dataset by Training VGG from Scratch", IJIST, Vol. 6 Issue. 4 pp 1691-1719, Oct 2024
Received| Sep 27, 2024 **Revised|** Oct 18, 2024 **Accepted|** Oct 22, 2024 **Published|** Oct 24, 2024.

Recognizing facial emotions is still a major obstacle in computer vision, particularly when dealing with complex datasets such as FER-2013. Advancements in deep learning have simplified the process of achieving high accuracy, yet obtaining high accuracy on the FER-2013 dataset with traditional methods remains challenging. The aim of this research is to analyze the effectiveness of Convolutional Neural Networks (CNNs) utilizing VGG16 and ResNet50 architectures through three different training methods: training from scratch, with transfer learning, and fine-tuning. Our research demonstrates that while training VGG16 from scratch achieved a validation accuracy of 67.23%, fine-tuning produced a slight reduction in performance at 64.80%. Conversely, ResNet50 struggled across all approaches, with the highest validation accuracy being only 54.69% when trained from scratch. We offer an in-depth analysis of these methodologies by utilizing confusion matrices, training durations, and accuracy measures to showcase the balance between computational expenses and model effectiveness. Our results indicate that, although transfer learning and fine-tuning offer rapid convergence, training from scratch may still be necessary for specialized feature learning in complex FER tasks. These results help in the continuous work of improving emotion recognition systems by maintaining a balance between accuracy and computational efficiency.

Keywords: Facial Emotion Recognition, Convolutional, Neural Network, FER-2013 Dataset, Pre-Trained Models, Fine Tuning, VGG16, Res Net 50.

Introduction:

Humans often use words as their primary means of communication, but they also use body language to express themselves and demonstrate what they mean. Recognizing the significance of gestures in conveying emotions is critical since they are an integral aspect of nonverbal communication among people. The exponential development of computer vision methods and deep learning algorithms has led to considerable FER and image classification breakthroughs in recent years [1]. Among the most prominent factors contributing to this boost are the appearance of large, high-quality, publicly available labeled datasets and the empowerment of parallel GPU computing, which enabled the transition from CPU-based to GPU-based training, thus allowing for significant acceleration in deep models' training [2].

Methods like Support Vector Machines, k-nearest Neighbors, and the Perceptron multi-model have been used by researchers to tackle FER in the past. Features such as Eigenfaces, Local Binary Patterns, face landmarks, and Texture were leveraged by these algorithms for information extraction. In terms of popularity and prevalence in FER, neural networks stand head and shoulders above the competition. Reasons why CNNs have lately found utility in deep learning include: Without requiring humans to manually extract features from raw picture data, they are straightforward to deploy and can provide adequate performance. Finding new features becomes much easier with deep learning. When working with big datasets, deep learning outperforms traditional machine learning. Without input data, deep learning algorithms cannot function. A mountain of data is needed to train deep learning systems correctly [3]. Data with a grid-like structure, like pictures, audio, or video, is best processed by convolutional neural networks (CNNs), a subset of deep neural networks. [4].

Among the most well-known and quickly developing areas of computer vision is the detection of facial emotions. When it comes to nonverbal communication, this study is invaluable, especially for the deaf population. Furthermore, it is crucial for the study of human behavior. It is useful for evaluating emotions and making diagnoses of mental disorders. Even the lie detector will be vulnerable to this technology. Because they convey so much nonverbal information, facial expressions play an essential role in human interaction. The ability to automatically recognize these expressions might be a game-changer in creating more realistic human-machine interactions. This is only one of many potential uses for it; others include AI, physiology, psychology, behavioral science, and medical care [5]. But there are computers that can use their emotion recognition capabilities to their advantage. Indeed, relationships in many sectors might be revolutionized by such a technology.

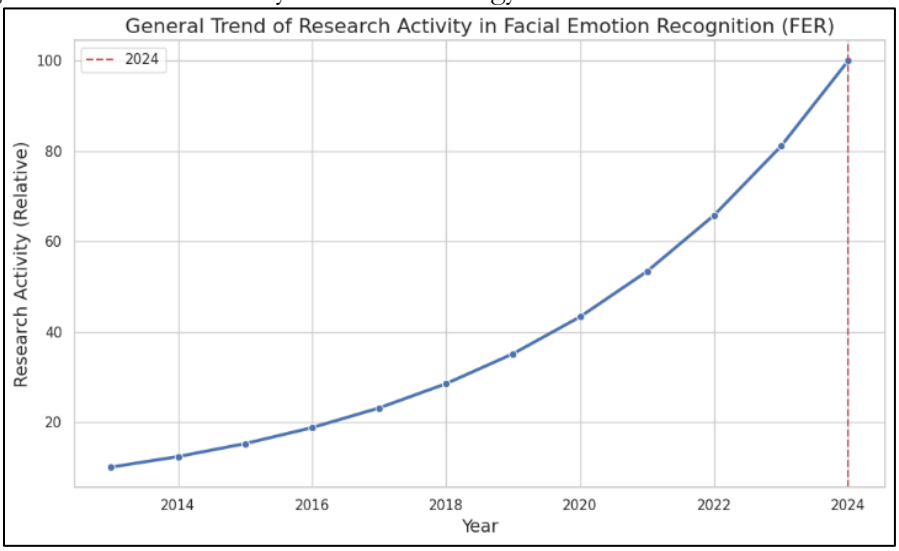


Figure 1. General Trend of Research Activity in Facial Emotion Recognition (FER) from 2013 to 2024

Many fields, including healthcare, social robotics, human-computer interfaces, and security, have taken an interest in it lately. The topic of how to get trustworthy FER in different contexts is, however, yet unanswered. Even with the advancements in deep learning and CNNs, it remains difficult to get high accuracy of validation on datasets like FER-2013. Additionally, FER-2013 is another famous benchmark dataset that was presented at the 2013 Worldwide Symposium on Machine Learning (ICML). It is a complicated yet suitable database for testing FER algorithms since it contains a huge amount of recorded facial expressions from various contexts. The present state of how people perform on that data set is impressive, despite the dataset's small size. Expected to be 65.5 % [6]. Figure 1 also depicts that the number of publications has been growing exponentially in the last decade which represents the overall research trend from 2013 to 2024.

The main issue related to the FER-2013 dataset is that it is difficult due to variations in lighting, pose, and occlusion. Earlier used machine learning algorithms such as Multi-layer Perceptron, k-nearest Neighbors, and Support Vector Machines have not been able to provide good accuracy on this dataset. Many of these methods rely on hand-crafted features which are often not sufficient to capture the fine differences in facial expressions. Therefore, even by using deep learning, it is not easy to obtain high accuracy on FER-2013. However, CNNs have simplified the process of feature extraction and outperformed traditional approaches, the nature of the FER-2013 dataset has been a challenge to many models, including class imbalance and the presence of subtle and complex expressions. In addition, achieving high accuracy in real-time applications is a problem in its right since it complicates the problem. The FER-2013 dataset is not as difficult as before, but there are still some problems that can be addressed only by using deep learning. In comparing different methods and in general when benchmarking we remember the previous work trained on this dataset.

This study seeks to address these challenges by exploring and comparing different training strategies on CNN architectures, specifically VGG16 and ResNet50, to enhance the accuracy and reliability of FER systems. Our goal is to determine the most effective emotion classification strategy by comparing three training strategies – starting from scratch, transferring learning, and fine-tuning. It is important because it addresses existing methods' shortcomings, such as class imbalance and real-time processing challenges, while also considering FER's ethical implications. By developing more reliable and fair emotion recognition technologies, this research will improve human-AI interaction in various applications.

As follows, the manuscript consists of the following sections: Section 2, which discusses the various studies relating to facial expression recognition and deep learning techniques. The dataset and the challenges encountered when recognizing emotions are described in sections 3 and 4. VGG16 and ResNet50 are used in this study as model architectures. The research methodology, including preprocessing, splitting, and visualizing data, is described in section 5. We present the results of our experiment and compare pre-trained and fine-tuned models in section 6. In Section 7 we discuss comparative analysis of VGG16 and ResNet50. Finally, we conclude by offering suggestions for further research in Section 8.

Novelty and Objectives of Study:

Although architectures such as VGG16 and ResNet50 have been employed in prior works, our work is different as it employs these architectures in the FER-2013 dataset under three different training approaches: training from scratch, transfer learning, and fine-tuning. This study provides a comprehensive analysis of their precision and computational efficiency, offering insights into their performance in both accuracy and computational cost, which has not been fully explored in earlier works. Additionally, challenges like class imbalance and subtle emotional expressions were addressed through data augmentation, class weighting, and fine-tuning techniques, making the models more generalizable to real-world emotion recognition tasks.

The objective is to build an accurate emotion classifier capable of enhancing the accuracy of the FER-2013 dataset while maintaining practical applicability for real-time applications including healthcare monitoring and adaptive learning systems.

Literature Review:

There has been an excess of research on FER using different methods in recent years. Important information on emotions from images has been extracted using traditional machine learning methods in combination with CNN models. As pointed out by Sarvakar, Ketan, et al., one of the challenges in emotion classification using deep neural networks is the need for high reliability and accuracy. This CNN with six convolutional layers, two maximal layers of pooling, and two fully connected layers was able to obtain a 60% accuracy rate; this rate shed light on the limitations of emotion classification due to inadequate data [7].

Using a convolutional neural network (CNN) on the FER2013 dataset, Zahara, Lulfiah, and colleagues achieved 65.97% accuracy in emotion identification using Raspberry Pi deep learning models. To enhance emotion recognition performance, they stressed the need for more sophisticated CNN architectures, larger amounts of training data, and superior hardware [8]. In this study, Xiang et al. explore the use of MTCNN (Multi-task Cascaded Con to recognize faces and identify facial features all at once. The authors emphasize the possibility for improvement in the often-overlooked synergy between these activities. By referencing seminal studies such as Ekman et al.'s 1971 research on six typical looks and the use of neural network models (CNNs), the article highlights the significance and development of Face Recognition (FER). They provide a novel method for face identification and FER by modifying the MTCNN, which was first proposed by Kiapeng Zhang et al. et al. Their experiments on the FER2013 dataset reach a validation accuracy of 60.7%, and they want to improve performance and precision in subsequent rounds by adding more layers and filters [9].

By using neural network models and the Viola-Jones algorithm, Ali, Khatun, Turzo, et al. investigated the identification of facial emotions. Using decision trees along with deep learning for emotion detection, their model outperformed human visual systems on the Kaggle tool dataset, demonstrating the significance of robust databases in the development of AI [10]. Using a convolutional neural network (CNN) structure on the FER2013 dataset, Mukhopadhyay et al. examined feelings in online classrooms. They were able to identify typical emotions like contentment and discontent with 65% accuracy, but they had a hard time with dataset mismatch and lighting/camera angle fluctuations [11].

Using the ResNet50 architecture, Gaddam et al. achieved 75.45% training accuracy and 54.56% test accuracy in face emotion identification on the FER2013 dataset. They noted that larger and more evenly distributed datasets were necessary for the development of better emotion recognition algorithms. Using the RGB-D-T dataset, Prasad and Chandana demonstrated a DenseNet-based method for thermal facial emotion identification that achieved an accuracy of 95.7 percent. Their model required tuning to enhance its performance as it was quicker than prior models like SSD and YOLOv3 [12]. Singh et al. developed a CNN-based face emotion recognition system and tested it on FER2013; the results showed an accuracy of 61.7% without pre-processing. They recommended further research into data pretreatment and network optimization, and they brought up the fact that dropout layers may help the model run better [13].

The authors Rakshith et al. have designed the ConvNet-3 model specifically for emotion recognition with a training accuracy of 88% and a validation accuracy of 61%. They also stressed the need to improve generalization and pointed out that overfitting is a problem when working with small datasets such as CK+48 [14]. Irmak et al. analyzed emotion detection from facial expressions using CNNs and achieved 70,62% training accuracy on the FER 2013 dataset. They emphasized different datasets and encouraged the researchers to study the ways to enhance bias and optimize the model [15].

The work conducted by Huang et al. on CNN-based facial emotion detection achieved 63.10 percent effectiveness on the FER2013 dataset, out of three FER datasets. Label mistakes and data unpredictability were among the issues they covered, drawing attention to the need for higher-quality data and more precise algorithms. Findings from studies using CNN, the case of Dens VGG16, and Mobile Net for emotion identification demonstrate the development of trustworthy emotion analysis technologies. These developments point to the possibility of future progress in this area, as they have far-reaching implications in areas such as healthcare, psychology, human-computer interaction, and safety [6–16].

This work is by Dhvanil Bhagat and others wherein implementing Efficient Net, Res Net, and VGG Net based three more pre-trained models (Tabel 1), we found that they were not up to the level of Deep CNN. We recommend DCNN as the best model. The top emotion recognition model that was suggested achieved an accuracy of 82.56% on the training data and 65.68% on the results of the validation data [16].

Table 1. Summary of previously reported accuracies for the fer-2013 dataset

Method	Val. Accuracy	Year
CNN [13]	61.70 %	2020
CNN [8]	65.97 %	2020
Dense Net [12]	54.56 %	2021
VGG16 [6]	63.10 %	2022
Vgg16 [17]	46.58 %	2024
VGGNET [16]	51.11 %	2024
ResNet50 [16]	54.67 %	2024
Efficient Net [16]	58.41 %	2024
Dense Net [16]	60.35 %	2024
DEEP CNN [16]	65.68 %	2024
VGG16 from scratch (This Work)	**67.23 %**	**2024**

Dataset and Challenges:

Dataset Description:

The data used in our study comes from the FER2013 dataset [18], The data consists of 48x48 pixel grayscale images of faces. The faces have been automatically registered so that the face is centered and occupies about the same amount of space in each image.

Table 2. The distribution of training and testing set of the fer-2013 dataset [18]

Category	Training Set	Testing Set
Happy	7,215	1774
Sad	4,830	1247
Fear	4,097	1024
Surprise	3,171	831
Neutral	4,965	1233
Angry	3,995	958
Disgust	436	111

In above Table 2, we provide an overview of how images were distributed across different emotion categories in both FER-2013 training and testing subsets. A total of 28,709 images are included in the training set, including 4,324 images labeled "Angry," 434 images labeled "Disgust," 3,207 images labeled "Fear," 7,091 images labeled "Happy," 4,309 images labeled "Sad," 2,485 images labeled "Surprise," and 5,254 images as "Neutral". A total of 3,589 images are available in the testing set for "Angry," 59 for "Disgust," 400 for "Fear," 870 for "Happy," 577 for "Sad," 435 for "Surprise," and 728 for "Neutral." In the dataset, happy and neutral emotions are way more prevalent than disgust and surprise.

Figure 2. General Trend of Research Activity in Facial Emotion Recognition (FER) from 2013 to 2024 (generated with matplotlib).

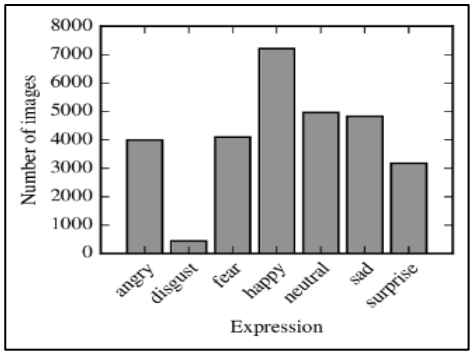
It can be seen from the data used in the study; that the following is a sample of the possible facial expressions as depicted in Figure 2. In each picture, there is a different person with different features and different emotions on the face, which makes it easier to describe the representations of these emotions. It has "Happy", "Sad", "Angry", "Surprise", "Fear", "Disgust", and "Neutral".

Dataset Challenges:

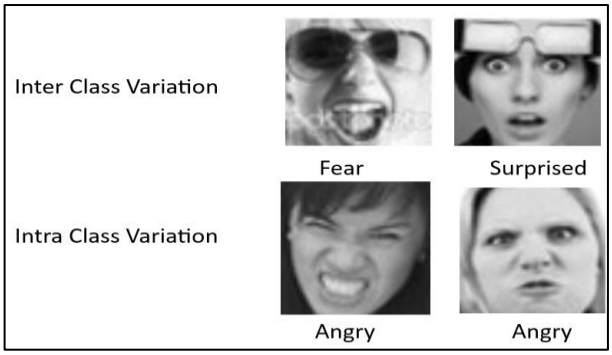
There are some challenges in the FER-2013 dataset which can lead to problems in the models for emotion recognition as shown in Figure 3.

- **Class Imbalance:** A highly skewed class distribution. For instance, while compiling the FER-2013 dataset, "Disgust" has been used less often than "Happy" and "Sad". Due to this, models can be trained on the examples of how to predict emotions while at the same time not being able to predict emotions with few examples.
- **Low Image Resolution:** The images in the dataset are only 48x48 pixels in size, the model may fail to recognize details of a person's face. This poor resolution can lead to an issue in the right identification of the emotion of the subject.
- **Inter-Class and Intra-Class Variations:** There are Interclass and intraclass variations in the dataset which are quite challenging. In some cases, it is even difficult to distinguish between different emotions in an interclass variation, for example, "Fear" and "Surprise" as the variations are quite similar. An intra-class variation can therefore be defined as the variation of the same emotion class in which an individual or a group may express the same emotion in different ways.
- **Diverse Facial Representations:** The dataset includes both real faces and cartoon faces, so the model should be able to learn from all sorts of facial images. Because of this diversity in facial representation, training models becomes difficult as they have to learn emotions in both kinds of pictures.

- **Occlusions and Pose Variations:** It means that the image can have occlusions, for example, the hands are on the face or in different poses. With such inconsistencies, the model must be able to handle them, and this makes the detection of emotions even harder.



(a) [19]



(b)

(c)

(d)

(e)

Figure 3. Preview of FER-2013 Dataset Challenges. **a.** class imbalance **b.** Inter-Class and Intra-Class Variation **c.** Occlusions and Pose Variations **d.** cartoon face **e.** 48x48 Low Image Resolution [18]

Model Architectures:

(VGG16 (Visual Geometry Group) Model):

One model that has seen extensive application for projects involving computer vision is a complex layered neural network (CNN) created by the Visual Geometric Group (VGG) and was evolved at the University of Oxford as shown in Figure 4 [20]. We can see that VGG16 consists of thirteen convolutional blocks & three completely linked layers. Every module of its five convolutional layer blocks has a max-pooling operation and two or three convolutional layers. After the convolutional blocks, the layers that are completely linked are added. The last layer, which includes a SoftMax Layer, is utilized for classification. Each convolutional layer in VGG16 maintains the input resolution thanks to 3x3 filters that have a speed of 1 and padding. For learning complicated characteristics, using tiny filters allows the network to have a big number and enhanced depth. To cut the spatial size of the mapping of features in half, VGG16 employs a 2x2 rectangle with a stride value of 2. This down sampling keeps the most critical details while reducing computing strain. Two fully connected layers, each with 4096 neurons, and four convolutional layers made up the initial ImageNet model. With almost 138 million parameters, VGG16 is a computationally demanding model because of its depth and completely linked layers.

(Res Net (Residual Neural Network) Model):

Res Net 50 is a deep convolutional neural network (CNN) architecture introduced in 2015 and is known as Residual Networks. They solved the vanishing gradient problem through the introduction of residual learning and allowed the networks to build deeper models. ResNet50 has a total of fifty layers; forty-nine layers are convolutional layers, and there is only one fully connected layer as shown in Figure 5 [21]. The model has five phases, and every phase encompasses several convolutional blocks. The use of remaining fragments which is known as

"skip connection" that enables the input to be added directly to the output after one or more layers are skipped is the basic innovation of ResNet50. For this reason, training deep networks becomes easier since the network learns identity mappings. Every residue block in ResNet50 has three convolutional layers, 1x1, 3x3, and the third 1x1 convolutional layer is a bottleneck structure. The network is made more efficient by the decrease of the number of more parameters that are due to the 1x1 convolution which both decreases and increases the number of channels. To learn properly in the extensive network, skip connections are needed and that results in the vanishing gradient problem. They allow a direct flow of gradients through the network and therefore allow optimization. Instead of fully connected layers, ResNet50 does not have them after the last convolutional layer, but it has global average pooling, which significantly decreases the number of parameters and the chances of overfitting. ResNet50 is deeper in design, and it uses residual connections to increase its speed but it has 25 million parameters which is much less than VGG16.

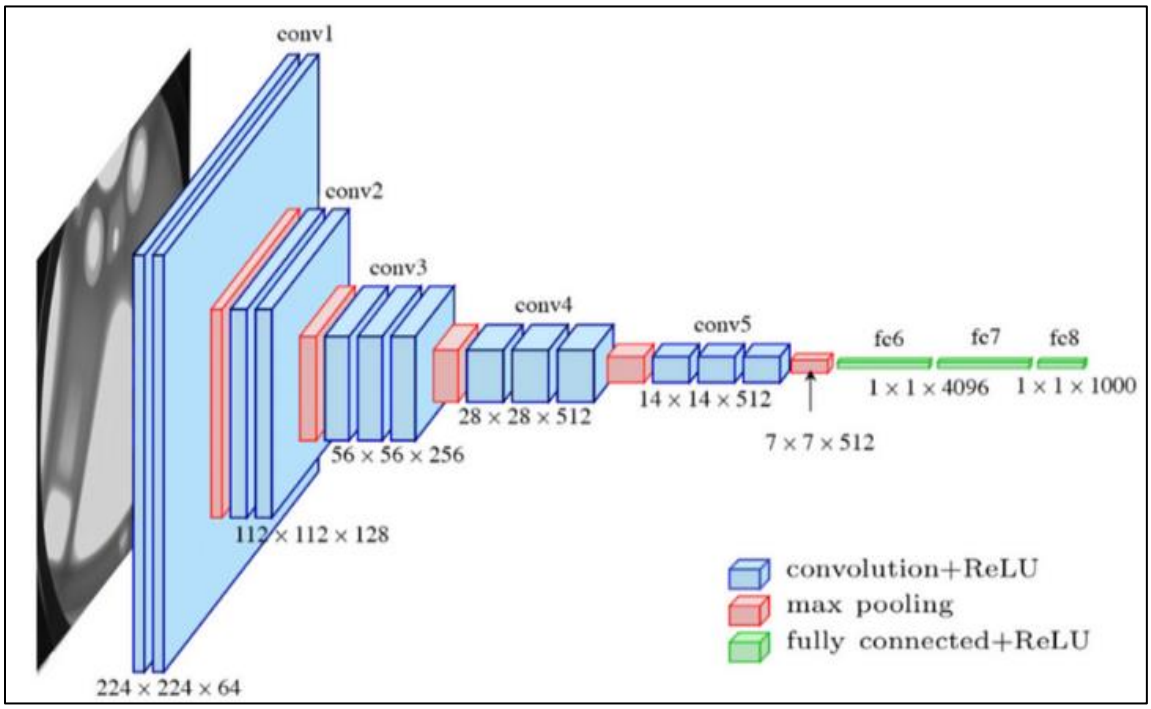


Figure 4. The architecture of VGG16 [20]

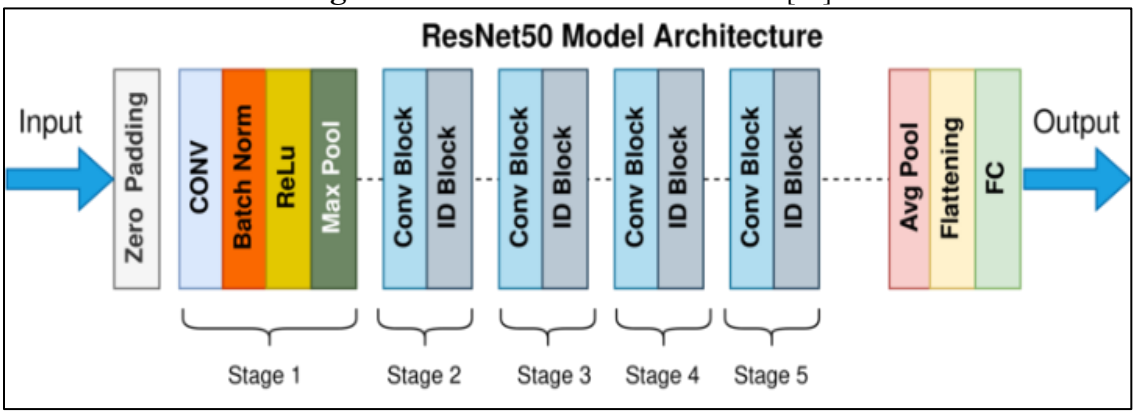


Figure 5. The architecture of ResNet50 [21]

Materials and Methods:

Training from scratch, transfer learning, and fine-tuning were the three methodologies explored on the FER-2013 dataset in relation to emotion detection utilizing the VGG16 and ResNet50 architectures. VGG16 & Res Net 50 models were trained only on the FER-2013

dataset and assigned random weights in the Train from the Scratch method. Because the models need to learn the features from the beginning, this approach was time-consuming, but it sought to learn the task-specific characteristics directly from the dataset. In contrast, the Transfer Learning method made use of ResNet50 and VGG16, two models that had already been trained on the massive ImageNet dataset. Using the characteristics learned from the ImageNet dataset, these models received education using the FER-2013 dataset. Because the models could employ the characteristics learned for emotion identification, convergence was quicker, and generalization was better. As a last point, the Fine-Tuning method included trainable layer weights for the FER-2013 emotion identification test. Emotion categorization benefited most from this method, which fell in between task-specific pre-training and feature learning.

The study contains many important metrics for each approach: To compare models that use the categorization of various emotions, the Confusion Matrix Training Duration for evaluating the model's computational demands, and the purpose of comparing the model's overall performance in the sets used for validation and training is to determine their validation accuracy and model accuracy. Metrics like this show how well each strategy is doing and how much efficiency each one has gained.

Data Preprocessing:

The FER2013 database will undergo preprocessing before CNN training to enhance the model's quality and performance for facial emotion detection. Using the Image Data Generator function, which is part of the Keras package, one may generate picture collections with enhanced data. Variations between classes in the dataset included both cartoonish and human faces, which added complexity to the operation. These differences were anticipated, and a template was built to be able to deal with them. The values for each pixel were divided by 255.0 to obtain the data for the band [0,1]. This was done to feed the mathematical model of the data and avoid memory concerns.

So that they would work with the VGG16 and ResNet50 systems that had been created on the ImageNet database, all the photos were reduced in size from 48x48 to 224x224 using bilinear interpolation as shown in Figure 6. Because your transfer learning algorithms accept inputs of varying sizes, we need to resize the image in the dataset, which is 48*48 pixels in size. (224x224).

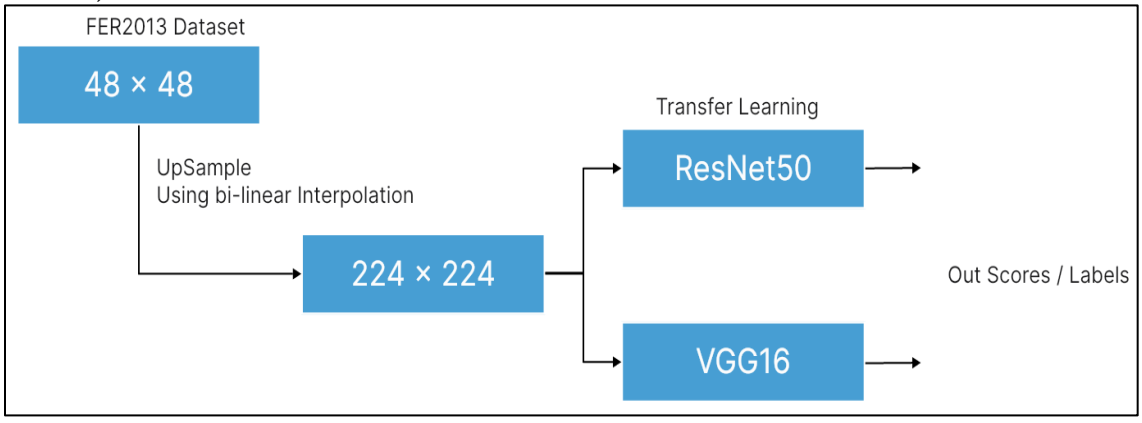


Figure 6. Bi-linear Interpolation

Training Data Preparation:

You have two methods to create training data:

- **Manual Loading:** Your initial code loads images into memory, preprocesses them, and creates NumPy arrays for training.
- **Image Data Generator:** This method uses a generator to load images on the fly, providing potential memory efficiency and augmentation capabilities.

The distribution of images based on each class is shown in Figure 7.

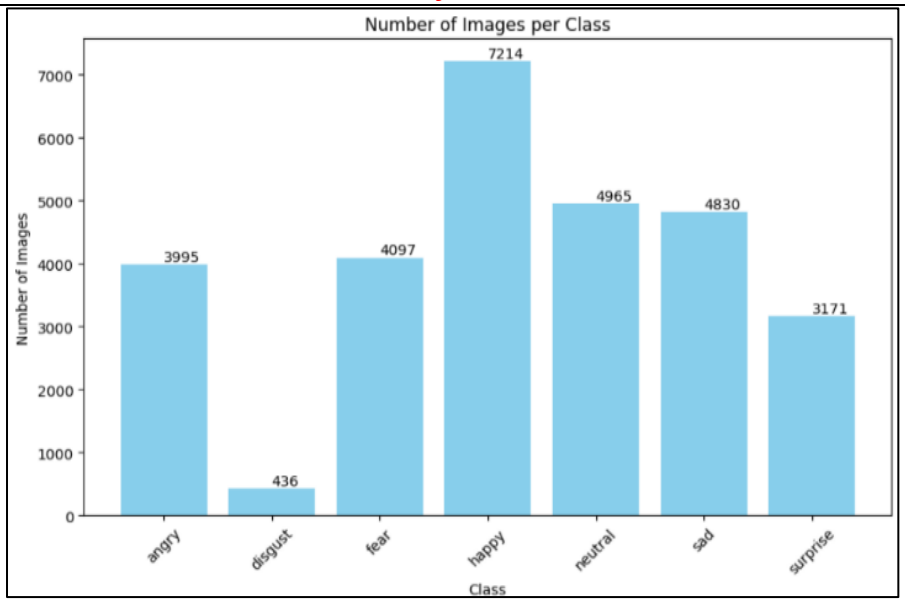


Figure 7. Class distribution in training data.

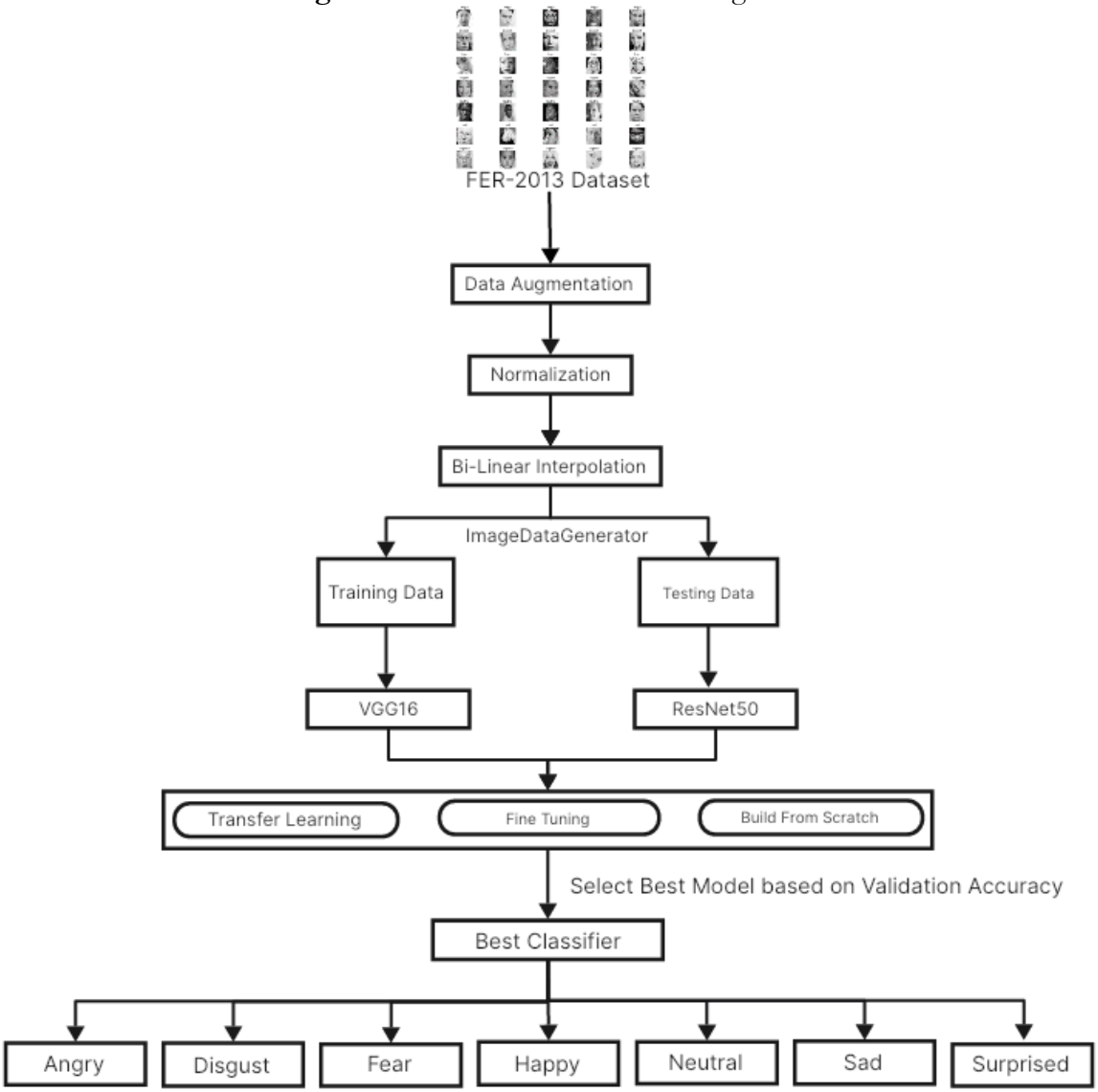


Figure 8. Our Proposed Methodology

Figure 8 summarizes the research process conveying the workflow of the process of data preprocessing, model training, and evaluation alongside performance comparison. It illustrates the three approaches of training – training from scratch, transfer learning, and fine-tuning with the FER-2013 set for facial emotion recognition.

Results and Discussion:

Experiment Setup:

The experiments were performed on a high-performance computer. Configuration with the following characteristics:

- **CPU:** AMD RYZEN 9 5900X
- **GPU:** NVIDIA GEFORCE RTX 4080 SUPER 16G
- VENTUS 3X OC
- **Memory:** 32 GB RAM

Pre-Trained Models:

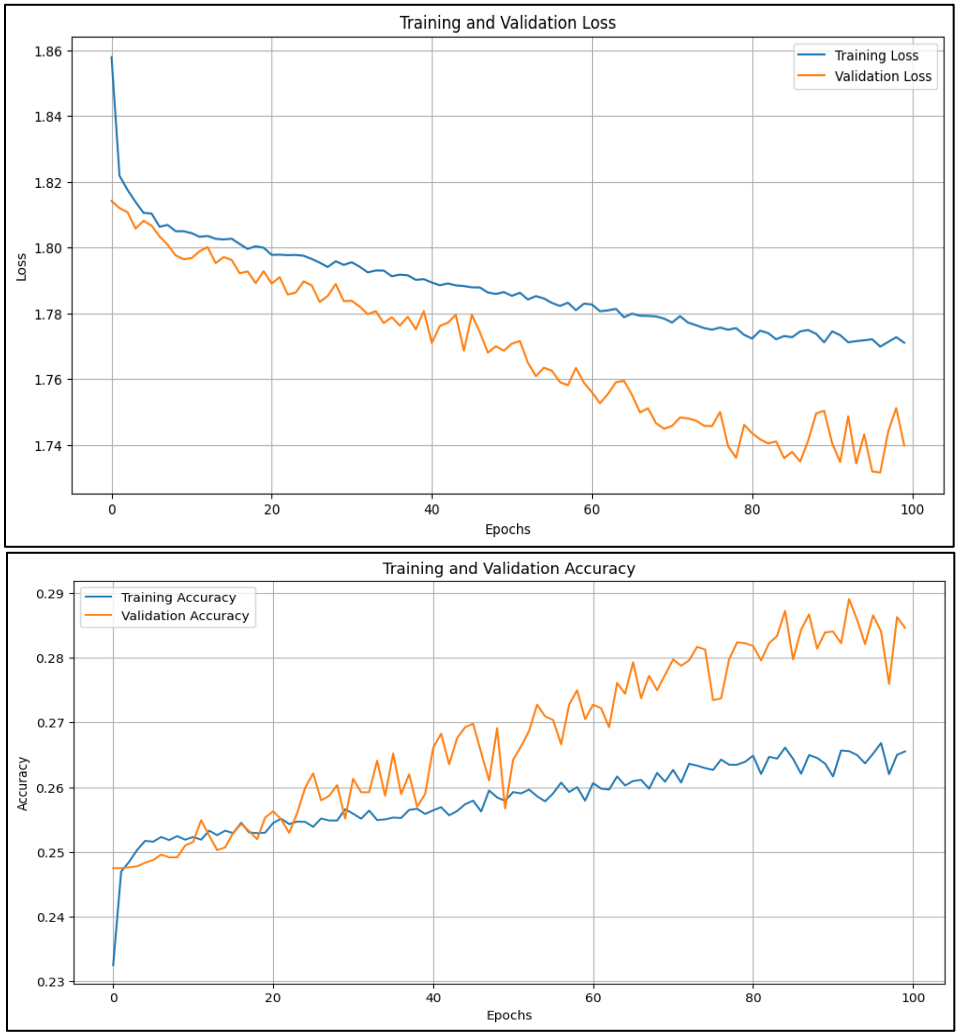


Figure 9. Performance Graphs of Res Net 50 with Pre-Trained Weights

Res Net 50 with Pre-Trained Weights:

For this, we additionally used pre-trained weights to refine the ResNet50 model, which improved the accuracy of the suggested emotion classification framework on the FER-2013 dataset. The dataset is structured into seven categories of emotions and consists of grayscale photos with dimensions of 48 by 48 pixels. The photos were changed to RGB format and resized to 224 × 224 pixels for normalization. Using the Image Data Generator class for horizontal flipping, rotating, altering its height and width, shearing, and zooming, and improving the

model's generalizability, we increased the variety of the training data. Additionally, the data used for validation was simply resized to maintain its original attributes and features.

Because it was developed for a different category task, the last layer was excluded from the ResNet50 model that was trained on the Image Net data set for this study. The features were down-sampled using an average pooling layer, a dense layer having ReLU activation parameters was applied for non-linearity, and a dropout layer was used for regularization. A dense layer using the SoftMax activation function provided the input probability for each of the seven emotions in the final layer.

It is a multiclass classification issue; the model was built using the Adam optimizer with an average rate of learning of 0.0001 and a loss function of categorical cross-entropy. In addition, we used techniques like early halting, which causes training to halt when validation accuracy exceeds 75%, and a lower learning rate on the plateau, which involves halving the learning rate if proof loss is not reduced.

There was a cap of 100 epochs for training, during which the instruction set was supplemented and the outcome of the model was evaluated using the verification data set. A number of metrics, including training duration, accuracy, and loss, were used to assess the method and the training itself. With a final validation accuracy of only around 28.46%, it's clear this feeling classification is a challenging task that requires more model tuning or maybe a new architecture.

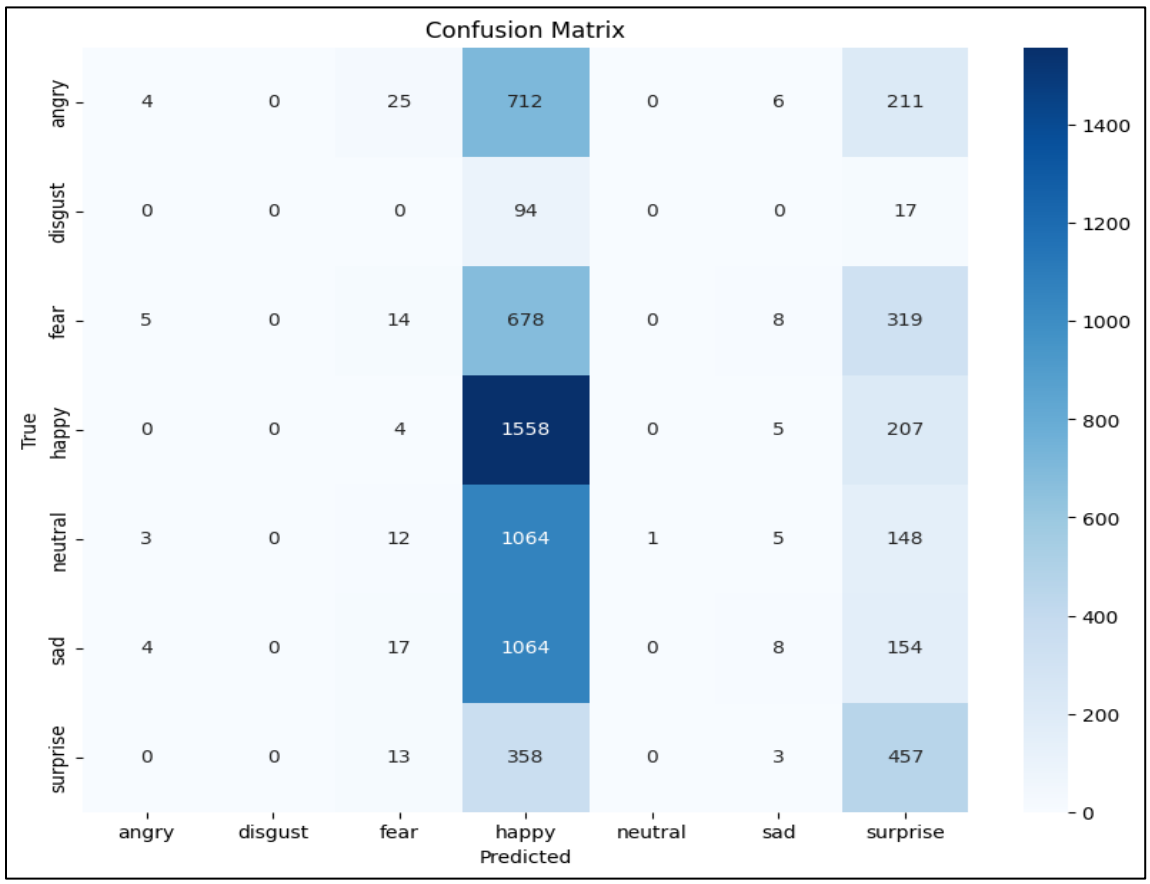


Figure 10. Confusion Matrix of ResNet50 with Pre-trained weights

Along with predicting results from the validation set, we calculated the matrix of misinformation of the model's classification outcomes from all seven classes to assess the model. In addition, we plotted the precision and loss curves for training and validation to evaluate trends in the model's learning. This method illustrates how to use a task-specific deep learning model for emotion identification and how challenging it is to do so on the FER-2013 database. Training

loss has been lowered from around 1.86 to 1.74, and when the sum of epochs rises, training loss and validation loss both drop, as seen in Figure 9. Validation accuracy was up 28.5% in the most recent iteration, rising from 0.24 to 0.285, as seen in the accuracy graph.

In Figure 10 the confusion matrix demonstrates the model's predictions across emotion categories, with "happy" having the highest correct predictions (1558). However, there's significant confusion between classes, particularly misclassifications of other emotions as "happy". The model shows signs of learning but has room for improvement, with relatively low overall accuracy and class imbalance issues evident in the confusion matrix.

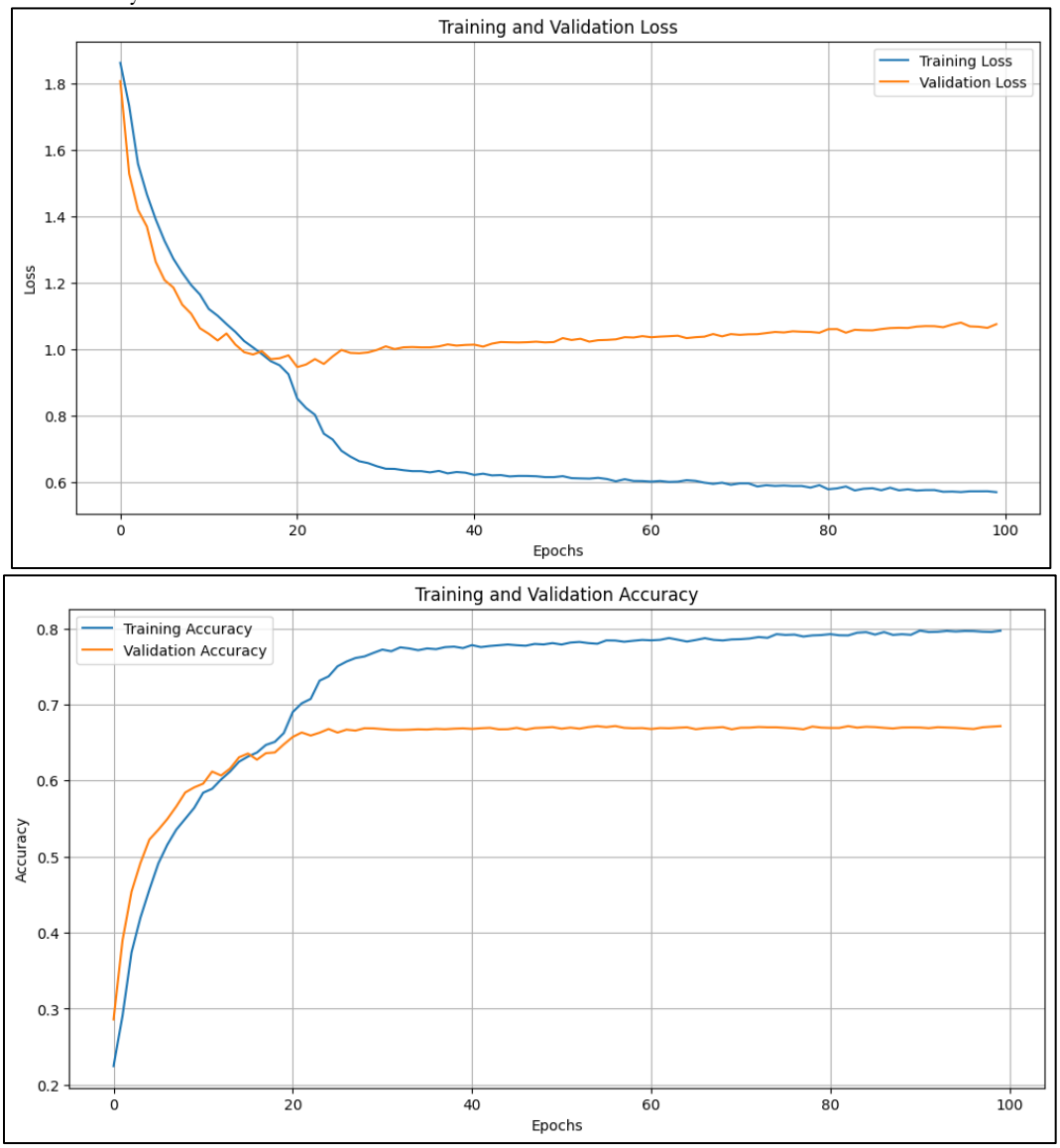


Figure 11. Performance Graph of VGG16 with Pre-Trained Weights

VGG16 with Pre-Trained Weights:

We used ImageNet's pre-trained weights in the VGG16 architecture to improve the emotion categorization. To prepare the FER-2013 dataset for input into VGG16, the 48x48 grayscale pictures were resized to 224x224 pixels and converted to RGB format. To build the model, many thick layers were superimposed on top of this previously trained VGG16 basis. To prevent overfitting, we inserted a dropout layer (rate=0.5) after flattening the convolutional base output, then layers that are completely linked with 64 neurons with ReLU activation. Before the final layer of output using the SoftMax activation function, there was another dense tier with 32 neurons with ReLU activation that was used to categorize the pictures into one of seven feeling

categories. The Adam optimizer was used to construct the model, with a rate of learning of 0.0001 and the loss function being categorical cross-entropy. We used two callbacks to guide our training: one to end training early when validation accuracy reached 75% and another to reduce the learning rate by half if rejection plateaued for two epochs in a row. The model underwent 100 iterations of training using data augmentation that included horizontal flipping, shearing, zooming, rotating, and alterations in width and height. About 16,304.71 seconds were required to complete the course. A validation loss of 1.0750 was attained, leading to a final validation accuracy of 67.18%. On the other hand, a loss of 0.5686 brought the training accuracy up to 79.69%. Insights into the model's performance among the seven emotion classes were supplied by the confusion matrix, which was constructed from the validation predictions. It highlighted areas of difficulty. Overfitting or a requirement for further modifications were among the tendencies shown by both the precision and loss curves shown during the training period. The experiment showed that VGG16 could be used for emotion categorization and also showed where the model might be improved to make it even better.

Figure 11 shows how a VGG16 with Pre-Trained Weights is performing as it trains over time. In the first plot, we see the training loss, represented by the blue line, which starts off high but drops significantly as the model learns. This decrease indicates that the model is improving its ability to predict the training data. However, the validation loss, shown by the orange line, levels off at a higher value, which suggests that the model isn't performing as well on unseen data. This could mean it is overfitting, or simply not generalizing well. In the second plot, we look at accuracy. The training accuracy (blue line) climbs steadily and approaches 80%, showing that the model is getting better at recognizing the training examples. On the other hand, the validation accuracy (orange line) rises more slowly, reaching around 60%. This gap indicates that while the model is mastering the training data, it struggles when faced with new data. Together, these plots emphasize the need to balance training success with the ability to generalize in new situations.

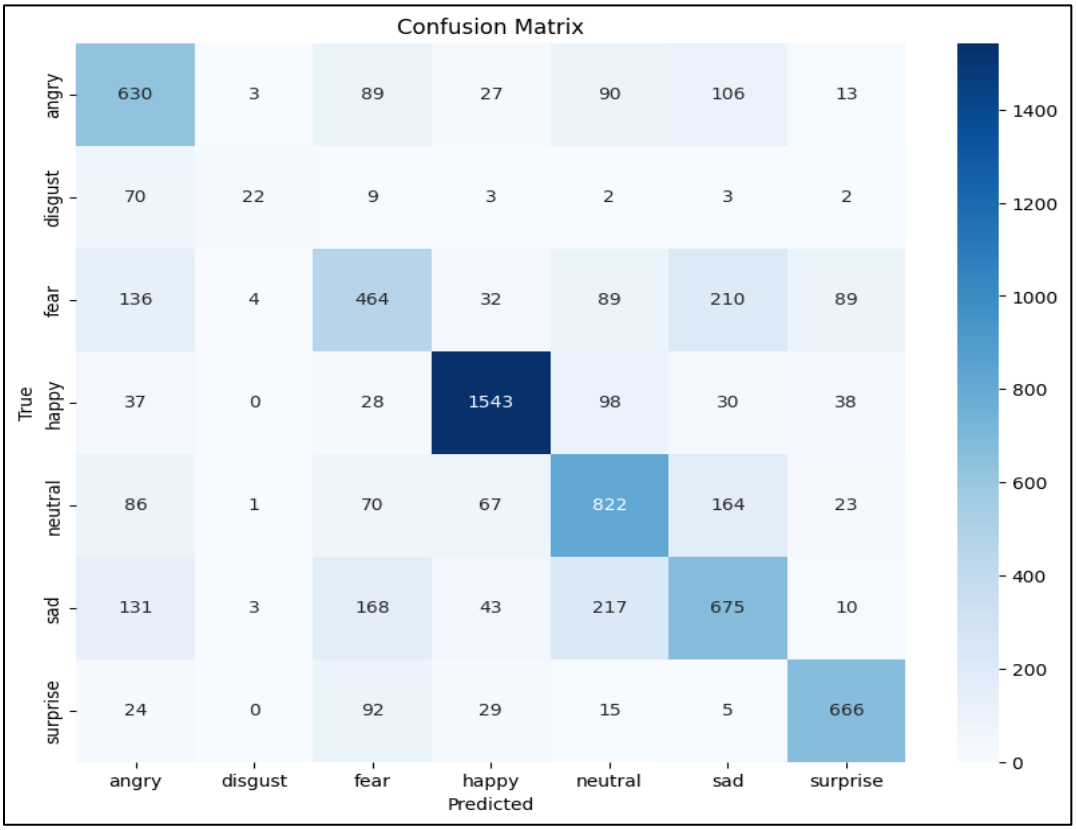


Figure 12. Confusion Matrix of VGG16 with pre-trained weights

The confusion matrix shows in Figure 12 that the model can identify "happy" emotions but has difficulty distinguishing between "fear" and "surprise," as well as "neutral" and "sad" emotions. Overall, the model performs well, but certain emotions can be better classified.

Fine-Tuning:

Res Net50 with Fine-Tuning:

This approach took advantage of ResNet50's robust feature extraction capabilities, which were initially trained on ImageNet. By utilizing its rich features learned from ImageNet, the ResNet50 architecture, known for its ability to handle deep networks effectively, has been adapted for emotion classification tasks. For this adaptation, we replaced the top layers of ResNet50 with a custom classification head. This new head included a flattening layer followed by dense layers with 64 and 32 units, utilizing ReLU activation functions. A dropout layer with a 50% rate was introduced to mitigate overfitting. The final layer employed SoftMax activation to classify the images into one of seven emotion categories.

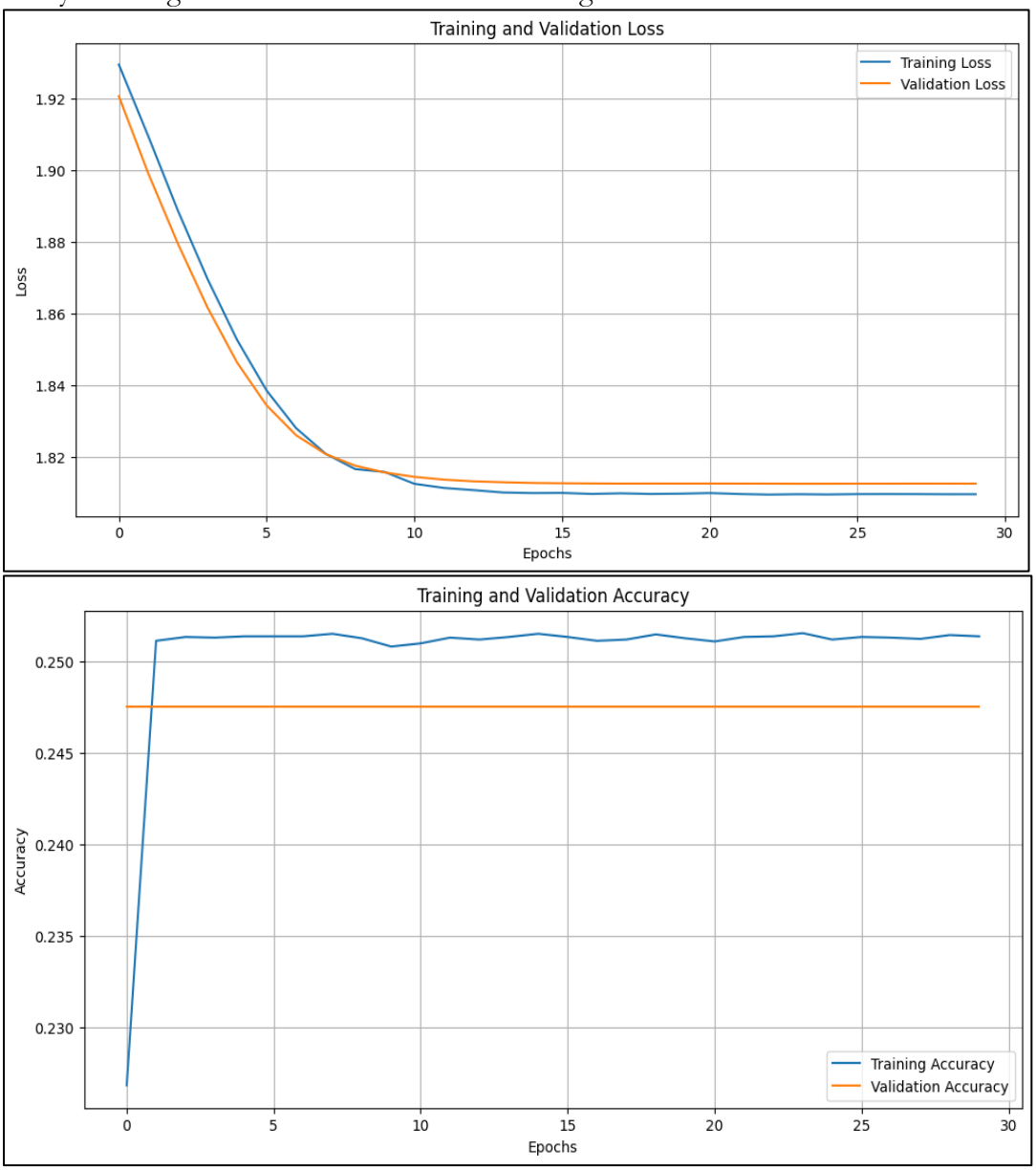


Figure 13. Performance Graph of ResNet50 with Fine-Tuning

Initially, the base layers of ResNet50 were kept frozen to retain the pre-trained features. During this phase, only the new layers were trained. In the later stage of fine-tuning, some deeper

layers of ResNet50 were unfrozen to allow minor adjustments to the pre-trained weights for better adaptation to the emotion classification task. The training utilized the Adam optimizer with a learning rate of 0.0001 and categorical cross-entropy as the loss function. Early stopping was employed to prevent overfitting, halting the training process when validation accuracy did not show significant improvement. The training was conducted over 30 epochs.

The fine-tuned model achieved a training accuracy of 25.14% and a validation accuracy of 24.75% after 30 epochs. These results highlight the challenges in adapting the model for emotion recognition, particularly with the nuanced differences in facial expressions within the FER-2013 dataset. Performance graphs showing accuracy and loss over epochs, along with a confusion matrix, are provided to visualize the model's classification performance across the seven emotion categories.

Figure 13 shows the training and validation loss decreasing rapidly in the first 10 epochs, then stabilizing around 1.41. Figure 13 displays the training accuracy plateauing at approximately 0.255 (25.5%), while the validation accuracy remains constant at about 0.245 (24.5%).

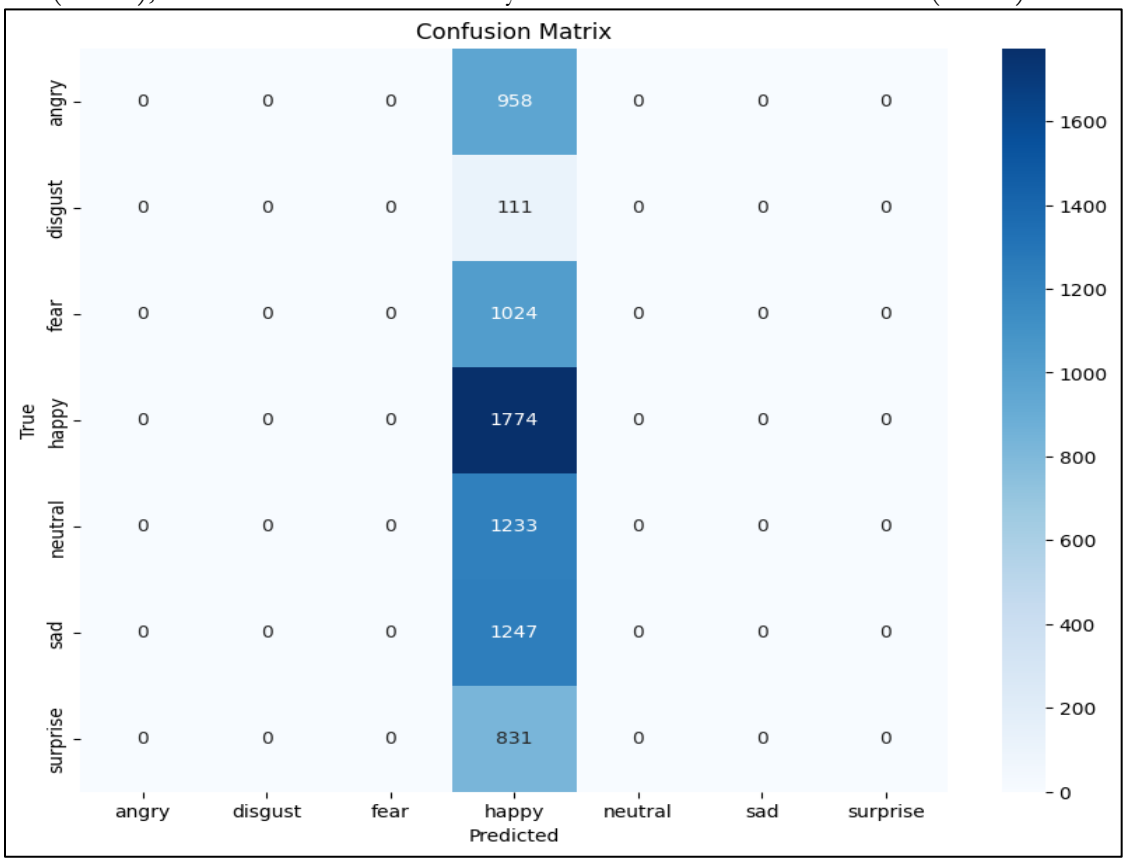


Figure 14. Confusion Matrix of ResNet50 with Fine-Tuning

In Figure 14. The confusion matrix in the third image reveals that the model predominantly predicts the "happy" emotion, with 1774 correct predictions for "happy" and misclassifications across other emotions. The model's accuracy is relatively low, as indicated by the accuracy graph, and it shows a strong bias towards the "happy" class, suggesting potential issues with class imbalance or model training.

VGG16 with Fine-Tuning:

The VGG16 model was customized for this purpose by adding a sequence of new layers to its basis, which was pre-trained on ImageNet. Before stabilizing learning using a batch normalization layer, we flattened the basis of convolutional output into a vector with just one dimension. After that, non-linearity was introduced via a thick layer including 64 neurons with ReLU activation. A dense layer in 32 neurons activated by ReLU and a dropout layer that a rate

of 0.5 were both used to reduce the likelihood of overfitting. In the last dense layer, seven neurons were used to categorize the pictures into seven different emotion categories using a SoftMax activation function. The model was trained using a categorical cross-entropy loss function and an Adam optimizer with a learning rate of 0.0001. We used early stopping-to-end training when validation precision reached 67% after a hundred years with an average batch number of 32. We kept an eye on the training process for 16,326.17 seconds. The model was tested on a validation set after training, leading to validation accuracy. of 46.50% and a validation loss of 160.38. These results, captured through confusion matrix analysis and accuracy/loss curves, indicated that the model's performance was not as high as anticipated, suggesting that further refinement or alternative approaches might be needed to improve classification accuracy.

In Figure 15 graphs illustrate the VGG16 model's performance during fine-tuning, focusing on loss and accuracy. The training loss graph shows a general decline with a significant spike around the 60th epoch, indicating potential overfitting or instability. Validation loss remains high, suggesting issues with generalization to unseen data. The accuracy graph shows a steady increase in training accuracy, reaching around 50%, but validation accuracy fluctuates, reflecting inconsistency in model performance. These results highlight the need for careful adjustment of hyperparameters and regularization to improve generalization and stability.

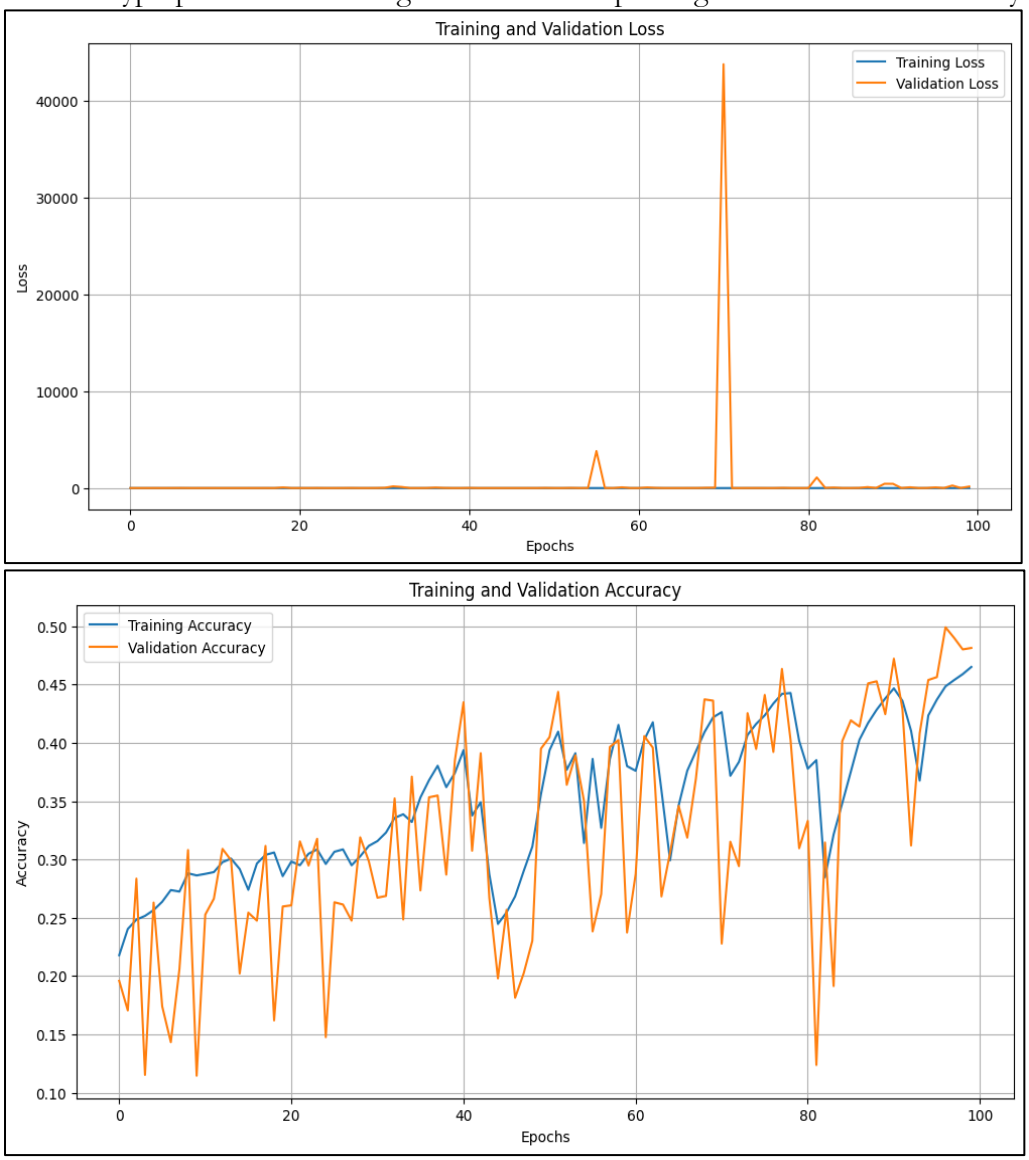


Figure 15. Performance Graph of VGG16 with Fine-Tuning

The confusion matrix shown in Figure 16 shows that the model excels in predicting the "happy" class but struggles with "disgust" and "fear," indicating challenges in distinguishing between some emotions.

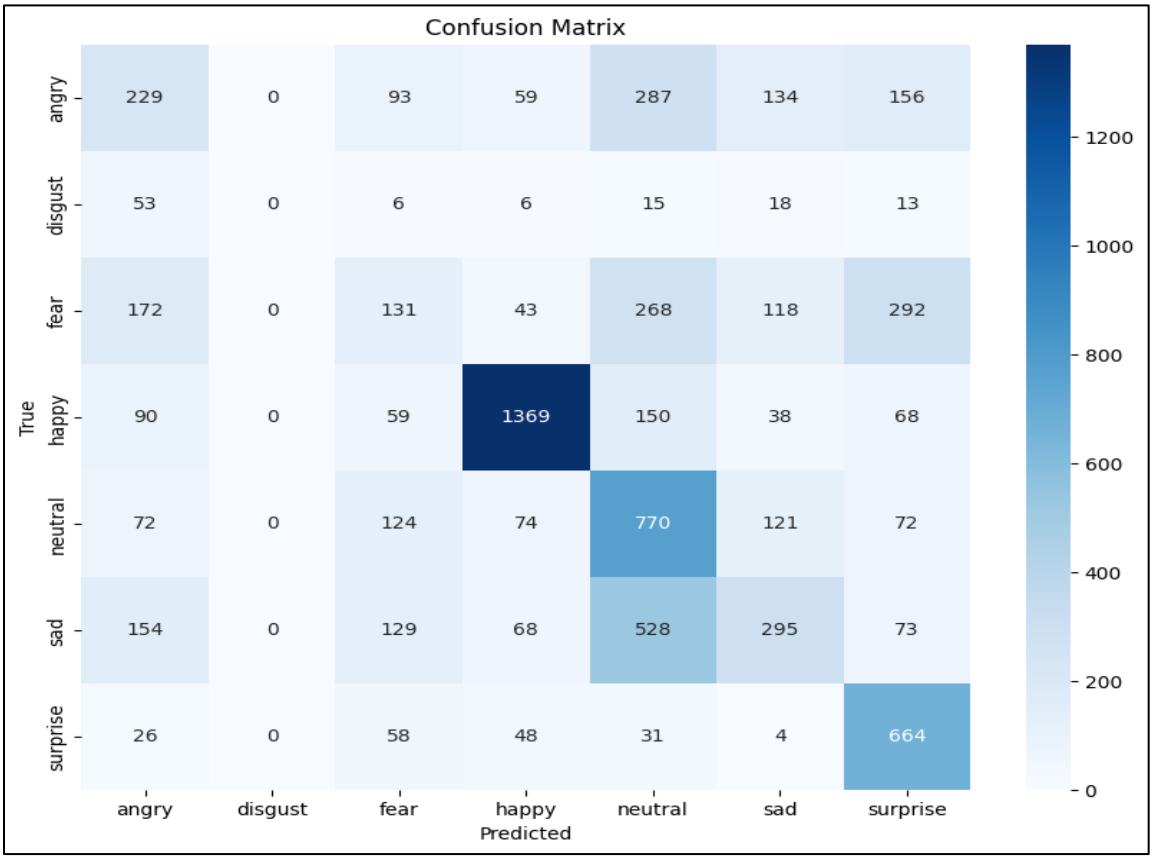


Figure 16. Confusion Matrix of VGG16 with Fine-Tuning

Training From Scratch:

Res Net 50 Trained from Scratch:

Res Net 50 architecture is a deep network with 50 layers, and it includes special skip connections. These connections help the model avoid common problems in deep networks, such as the vanishing gradient issue, by allowing the network to "skip" over layers and pass information directly through them. This structure helps the model learn more effectively.

To adapt the ResNet50 model for our emotion classification task, we have added several layers of our own. First, we utilized a global average pooling layer. This layer was applied to reduce the number of parameters and, therefore, the data was averaged across dimensions to guide the model to disregard unimportant characteristics and to prevent overfitting. Following that, we added another dense (fully connected) layer with 128 neurons and a ReLU activation function. It is also known as the learning layer and its function is to learn more complex patterns in the data. After that, we incorporated a dropout layer, which during training, sets half of the neurons off to prevent overtraining of the model. We then added another dense layer with 64 units The learned features are then passed through an additional fully connected layer with ReLU activation to enhance the feature learning. Finally, the output layer used the softmax function to predict the images to one of the seven emotions of the FER-2013 dataset. (e.g., anger, happiness, sadness, etc.).

Data augmentation was done to the photos during training. Randomly rotating, moving, or zooming the photos is one way to improve the model's ability to generalize to new data. To update the model's weights, we used the optimization algorithm Adam, which has a tiny learning rate of 0.0001. In addition, we used early stopping to prevent overfitting. Early stopping stops

training the model when it fails to improve on the set of validation data. It took about 24 minutes, or 1,445 seconds, to complete the course. The model achieved 56.18% training accuracy and 54.69% validation accuracy after 100 epochs. The model became better as it trained, but the confusion matrix revealed that it had trouble telling certain emotions apart. This could be because of the small details in facial expressions, particularly in grayscale photos. Training graphs display the model's accuracy and loss as a function of time. To help you see how well the model identified each emotion, we have included the confusion matrix. Figure 17 demonstrates that the model is instruction and generalizing as the losses from training and validation decrease over 100 epochs. Training and accuracy of validation increases, peaking at around 55% at the end of the era.

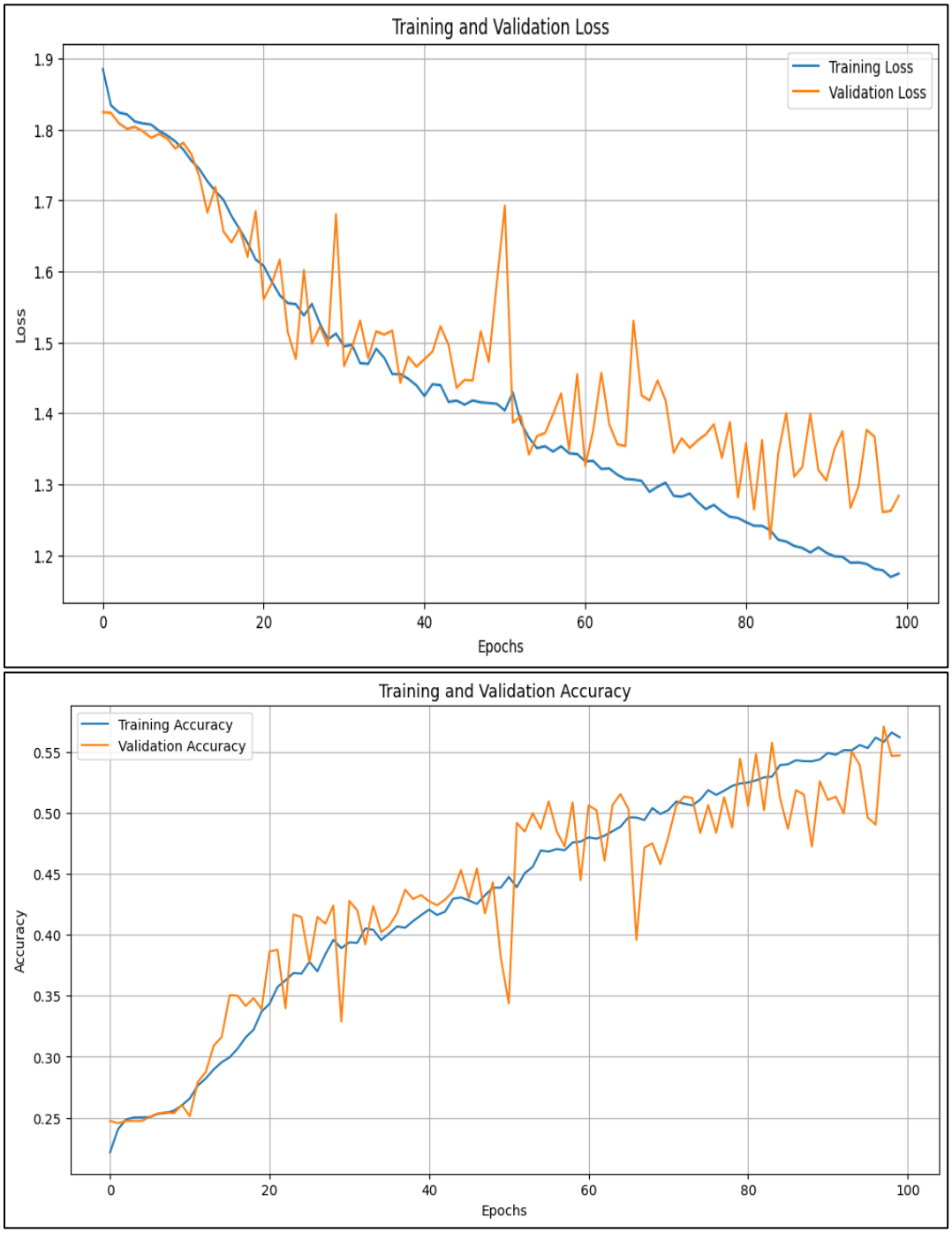


Figure 17. Performance Graph of ResNet50 from Scratch.

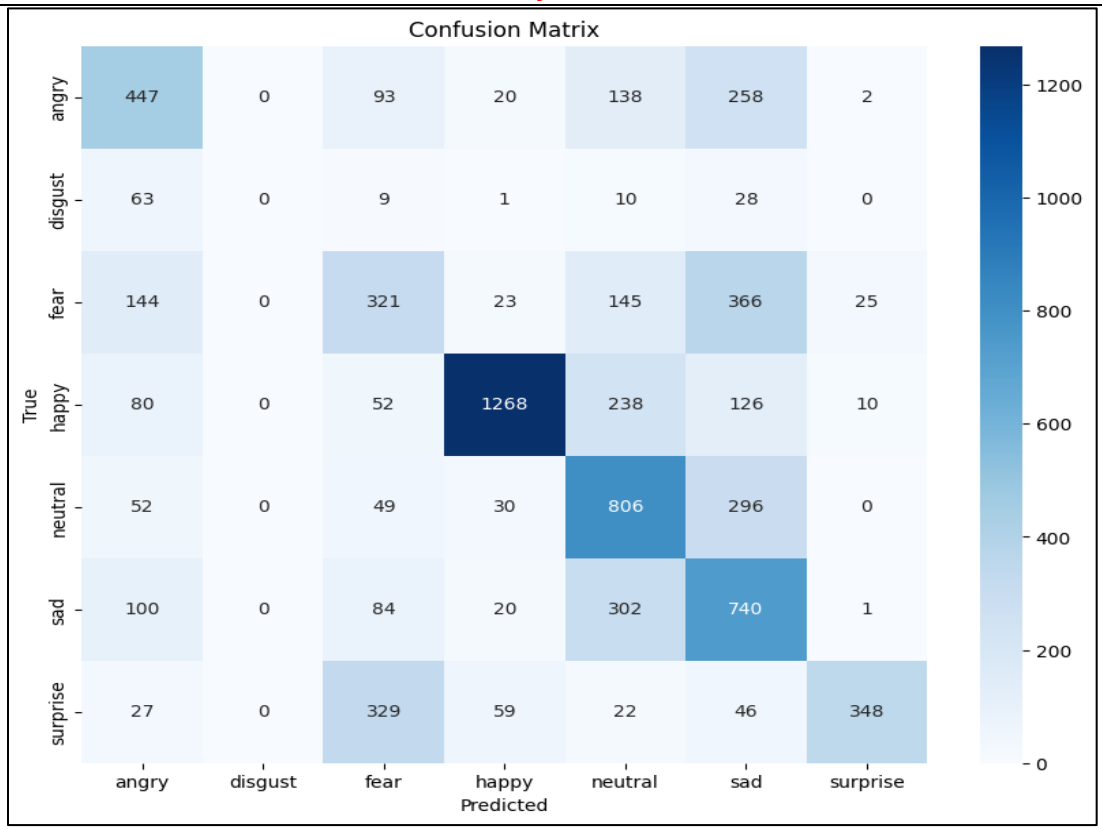


Figure 18. Confusion Matrix of ResNet50 from Scratch

Figure 18 presents a confusion matrix for the emotion classification task, revealing the model's predictions across 7 emotion categories. The diagonal elements show correct predictions, with "happy" having the highest accuracy. There's some confusion between certain emotion pairs, like "sad" and "neutral". Together, these metrics offer a comprehensive view of the model's learning progress, overall accuracy, and specific classification performance for each emotion category.

VGG16 Trained from Scratch:

The proposed model VGG16, model built from scratch and trained to classify facial emotions from grayscale images. The model architecture that was used had no pre-trained weights, several convolutional layers, and fully connected layers. Table 3 shows the vgg16 architecture. In training, rotation, width and height shifts, shear, zoom, and horizontal flips were applied as data augmentation to improve the model generalization. The augmentation was performed during training using a callback that stopped training once the model achieved more than 67% accuracy on the validation set. This callback was useful in avoiding overfitting the model and making sure that training does not go beyond the optimal point. The last model in the present paper reached the validation accuracy of 67.08 percent which may show better results than the models discussed in the paper. The training was performed 100 epochs, but the model was trained 78 epochs due to early stopping when the threshold of validation accuracy was achieved.

With reference to the performance assessment, the confusion matrix was deduced and ROC curves for each class were plotted. The confusion matrix offered a clear impression of how the model was performing in the classification and where it was good or bad. The ROC curves and the AUC values of every class provided information on how well the model could classify the emotions. For the model performance, the VGG16 model provided good accuracy with 67.08% validation accuracy and was better than the models used in this study. The time

taken for the training of the model was 1218.66 sec which is the time complexity for training the model with the chosen architecture and hyperparameters.

The training and validation accuracy is presented in Figure 19 together with the loss over 75 epochs. The training accuracy increases and fluctuates at approximately 72% while validation accuracy fluctuates at approximately 67%. The training loss is still decreasing; however, the validation loss is still almost 1, which means that there is no way to improve the model in general. These trends indicate learning and imply that the challenge of attaining high performance on the FER-2013 dataset remains high.

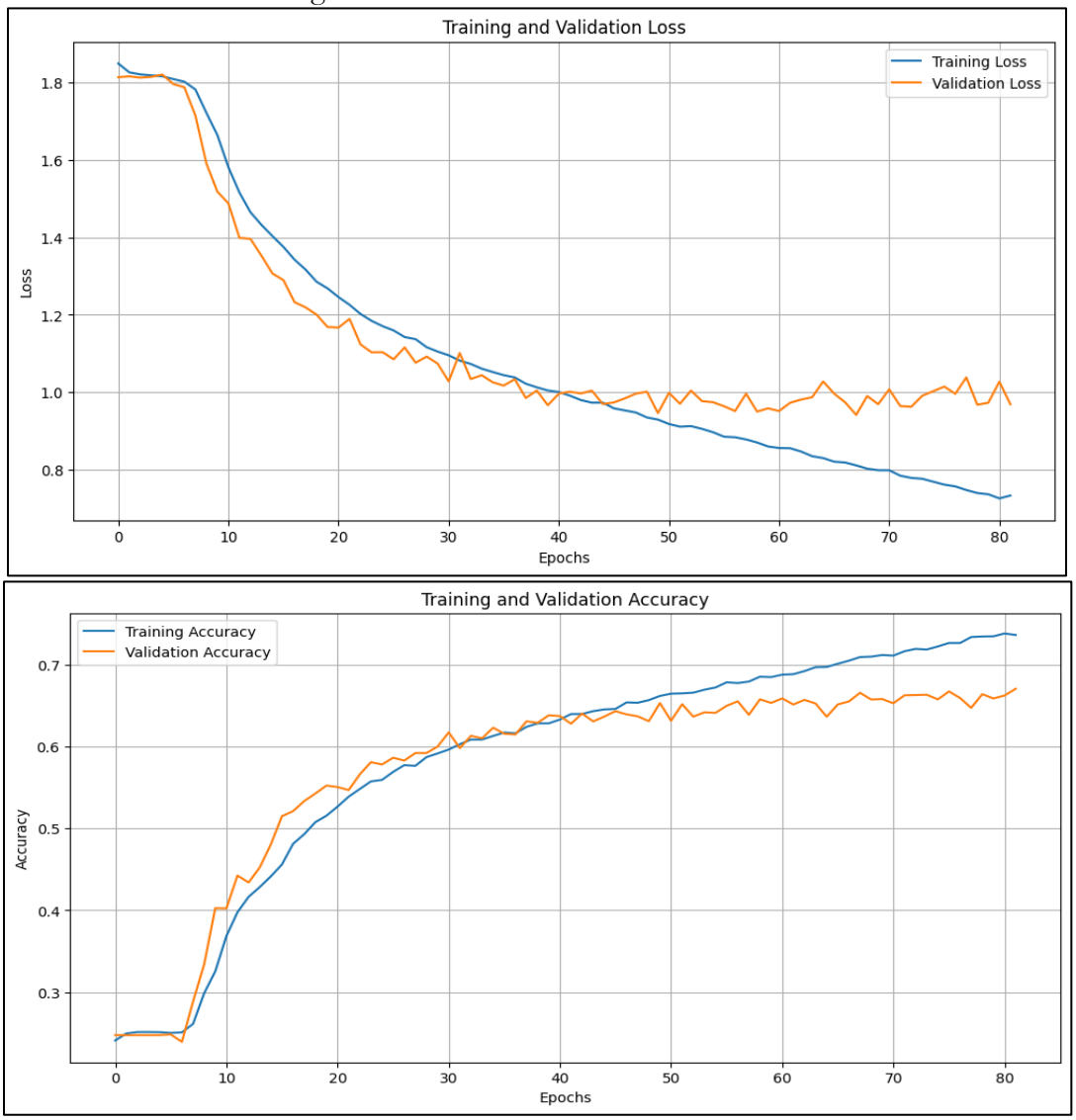


Figure 19. Performance Graph of VGG16 from Scratch

Table 3. Number of parameters for the Proposed VGG16 model

Layer Type	Output shape	Number of Parameters
Input layer	(none, 48, 48, 1)	0
Conv2d (block1_conv1)	(none, 48, 48, 64)	640
Conv2d (block1_conv2)	(none, 48, 48, 64)	36,928
Maxpooling2d (block1_pool)	(none, 24, 24, 64)	0
Conv2d (block2_conv1)	(none, 24, 24, 128)	73,856
Conv2d (block2_conv2)	(none, 24, 24, 128)	147,584
Maxpooling2d (block2_pool)	(none, 12, 12, 128)	0

Conv2d (block3_conv1)	(none, 12, 12, 256)	295,168
Conv2d (block3_conv2)	(none, 12, 12, 256)	590,080
Conv2d (block3_conv3)	(none, 12, 12, 256)	590,080
Flatten	(none, 4608)	0
Dense (64 units)	(none, 64)	262,144
Dropout (0.5)	(none, 64)	0
Dense (32 units)	(none, 32)	2,080
Dense (7 units)	(none, 7)	231
Conv2d (block2_conv1)	(none, 24, 24, 128)	73,856

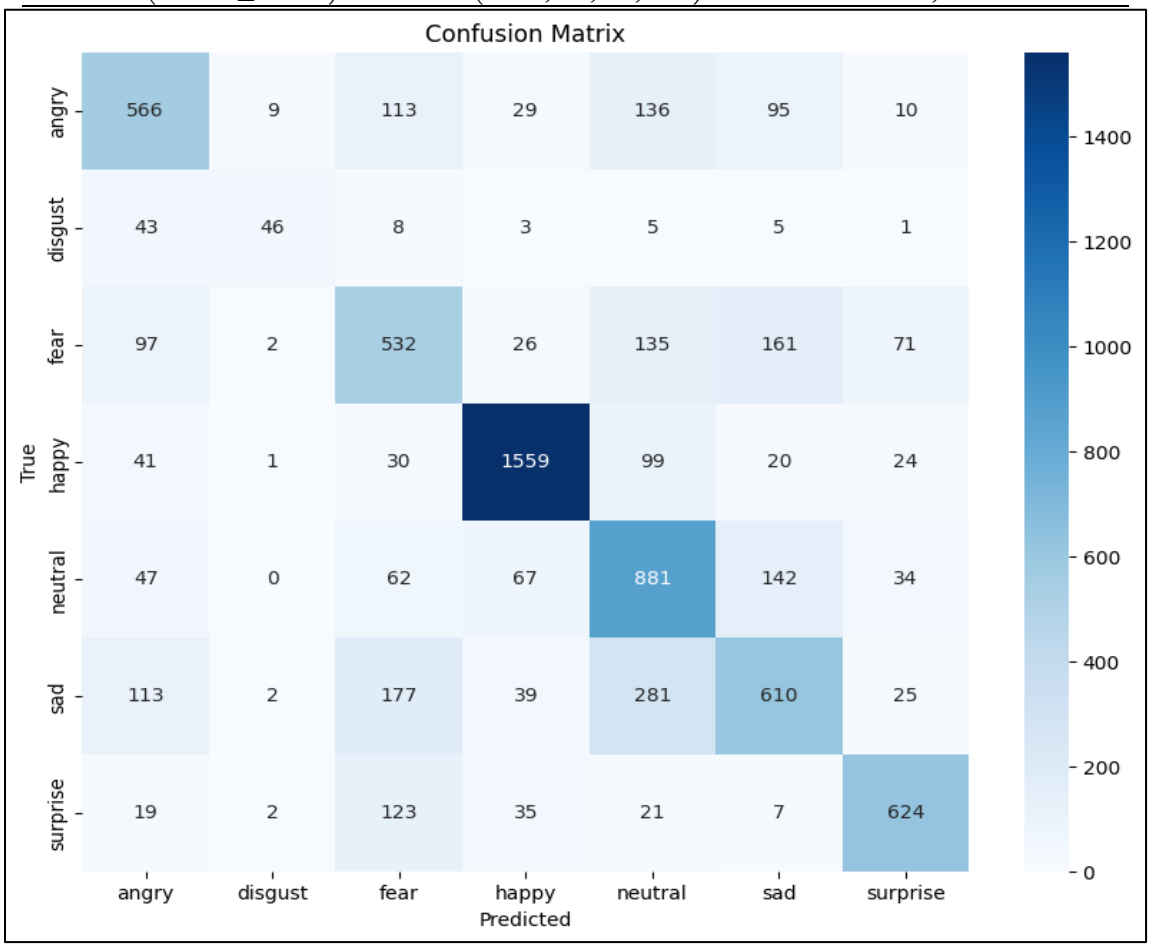


Figure 20. Confusion Matrix of VGG16 from Scratch

In Figure 20, the confusion matrix highlights the VGG16 model's strong performance in recognizing "happy" and "neutral" emotions but reveals struggles with subtle emotions like "disgust" and "fear," which are often misclassified. This reflects the inherent complexity and inter-class variation present in the FER-2013 dataset. The relevance of the confusion matrix lies in providing a detailed understanding of how well the model distinguishes between different emotions, beyond just overall accuracy. It helps identify specific areas where the model excels or needs improvement, offering insights into potential refinements for better real-world performance.

Figure 21 shows the Classification of Emotions Receiver Operating Characteristic (ROC) curves for each class, which includes anger, disgust, fear, happiness, neutrality, sadness, and surprise. In each scenario, the figure shows how well the model can differentiate between the specific emotion and all the other possible emotions.

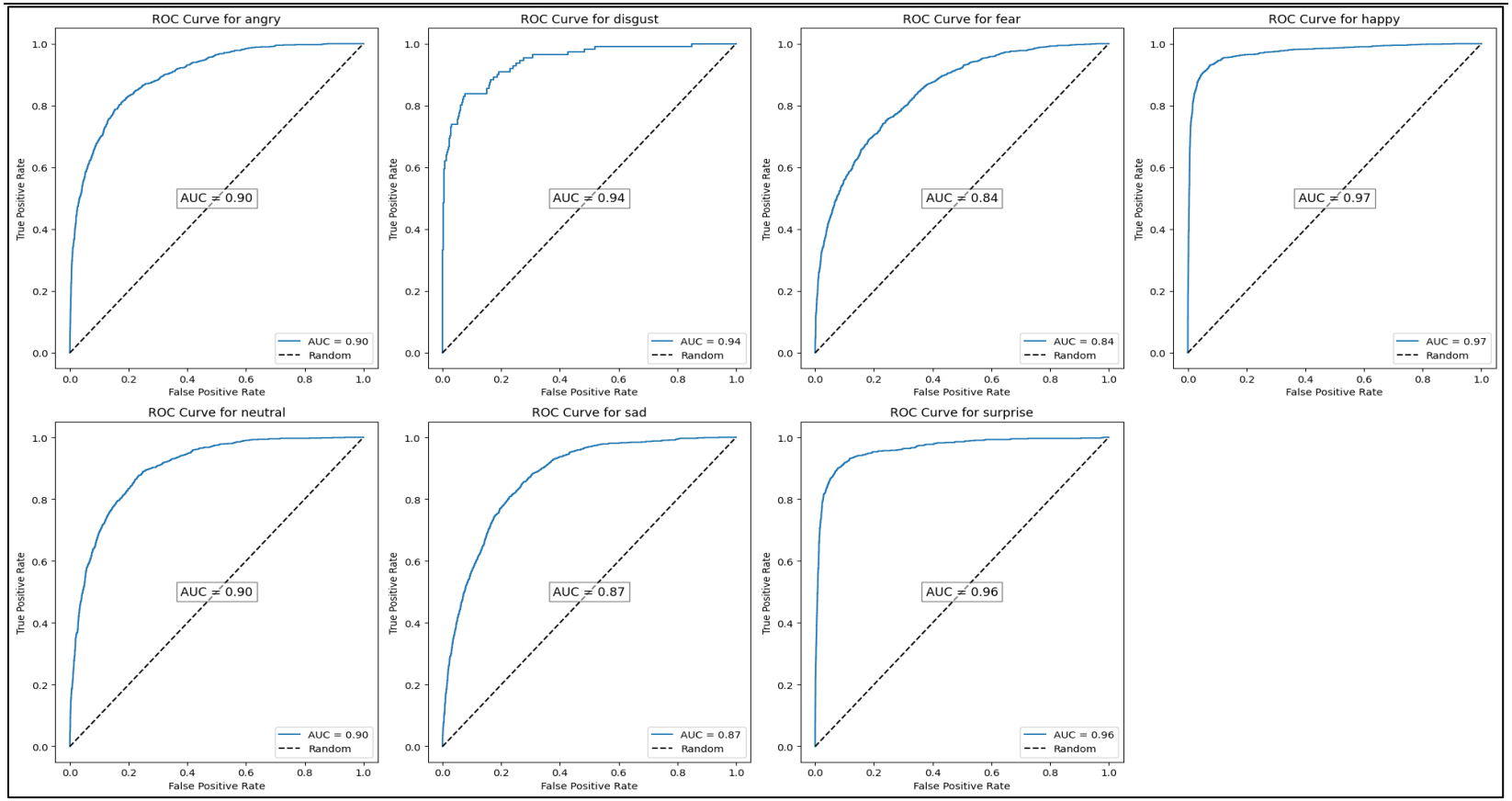


Figure 21. ROC curves for each class.

Discussion:

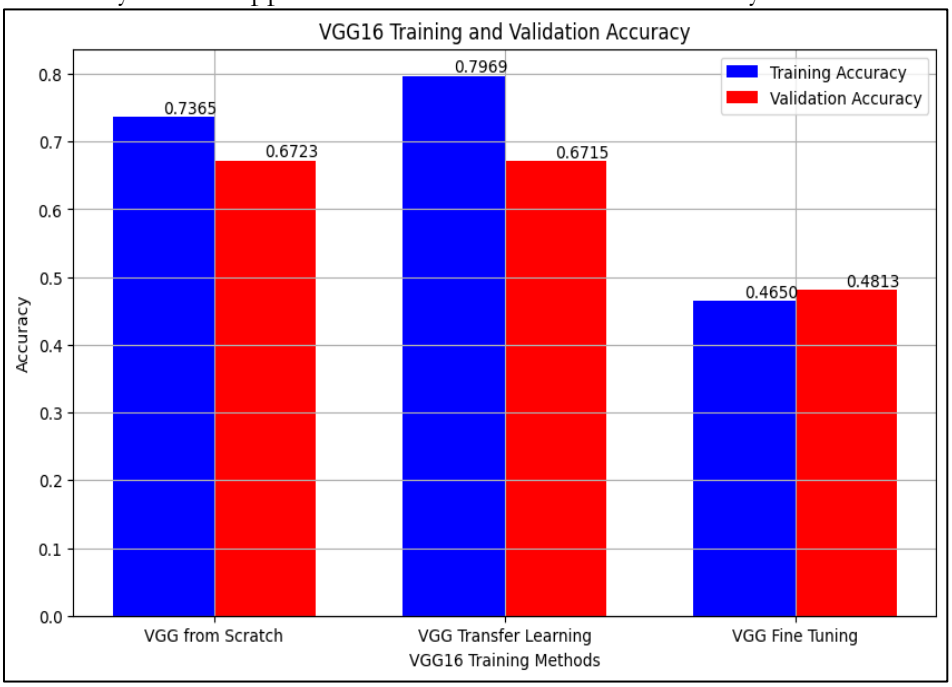
Table 4 also shows the previously reported classification accuracies on FER2013. It is observed that all the reported methods are better than the estimated human performance. (~ 65.5 %).

Table 4. Comparison table of evaluated models with existing literature on the FER2013 dataset

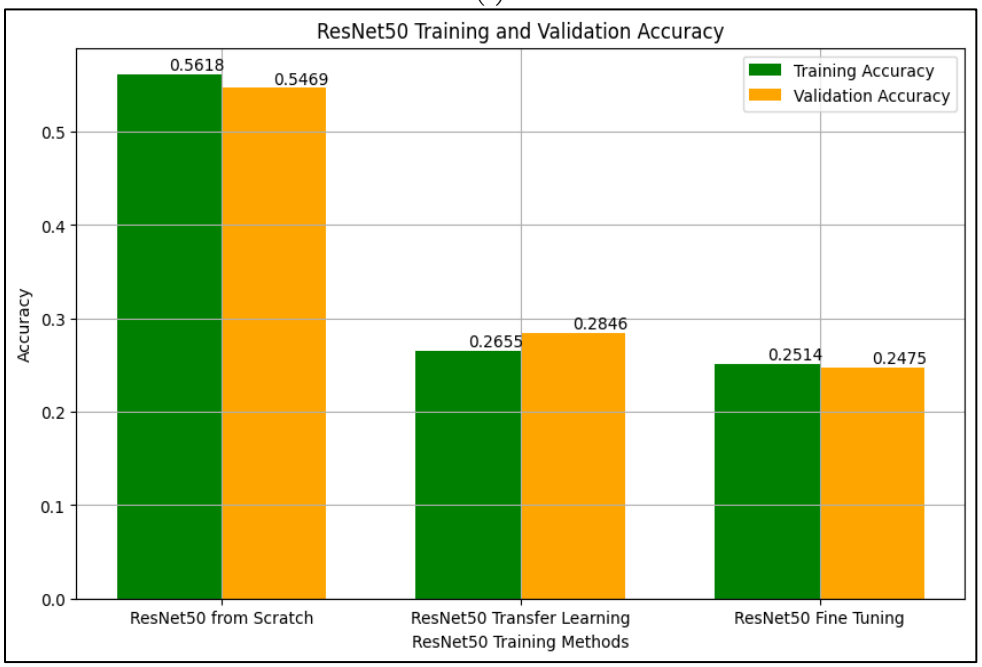
Method	Val. Accuracy	Year
VGGNET [16]	51.11 %	2024
ResNet50 [16]	54.67 %	2024
Efficient Net [16]	58.41 %	2024
DEEP CNN [16]	65.68 %	2024
VGG from scratch (This Work)	**67.23 %**	**2024**

Comparison of VGG16 and Res Net 50:

The two sets of bar graphs compare the performance of two different convolutional neural network architectures—VGG16 and ResNet50—across three training approaches: The three approaches of transfer learning are training from scratch, transfer learning, and fine-tuning. Each graph presents training and validation accuracy, which allows for making conclusions about the efficiency of each approach in terms of the model's accuracy.



(a)



(b)

Figure 22. a. VGG16 Training and validation accuracy **b.** ResNet50 Training and Validation accuracy

In addition to accuracy, the confusion matrices of both models provide deeper insights into the classification capabilities of each architecture. In terms of accuracy, both models had

fairly reasonable performances in emotions such as happy and neutral, yet in fear and disgust, the classification of each emotional set was quite off track. This difficulty further proves that the FER-2013 dataset is complex because of the fine differences in emotional gestures and due to the class bias.

For the VGG16 graph, we observe that the model trained from scratch achieves approximately 73.7% training accuracy and 67.2% validation accuracy. From these results, it would be seen that the model has a reasonable capacity to learn from the training data but not a very great capacity to generalize from the training data to other unseen data as depicted in the training and validation accuracies. The transfer learning approach improves the training accuracy as well as the validation accuracy to 79.7% and 67.1% respectively. This method uses weights from another model which was trained on a larger data set and this helps the VGG16 model to learn from the data without overfitting as would be the case if training from scratch.

However, the fine-tuning approach has lower accuracy with training and validation accuracy of 46.5% and 48.1% respectively. This decline may have been caused by overfitting where the model is trained to fit the training data at the expense of validation data by fine-tuning. In general, VGG16 results prove that transfer learning is the best approach as it gives high training accuracy with good validation accuracy.

On the other hand, the graph of ResNet50 shows a quite different picture. The model trained from scratch provides a training accuracy of approximately 56.2% and a validation accuracy of approximately 54.7%. Although these numbers are somewhat lower than in VGG16, they demonstrate that the ratio of training to validation accuracy is slightly better for ResNet50, which means that, despite the overall lower accuracy, ResNet50 might have a slightly better generalization. The transfer learning approach of ResNet50 reduces the training accuracy to about 26.5% and validation accuracy to 28.5%. This considerable reduction might be attributed to the point that the initial weights were not very suitable for the given data set in this case and hence the model is not as effective.

The fine-tuning method for ResNet50 also provides a low performance with a training accuracy of about 25.1% and a validation accuracy of 24.8%. This also provides more evidence to the idea that fine-tuning in this context does not enhance the performance, because the model does not optimize for the new data after pre-training.

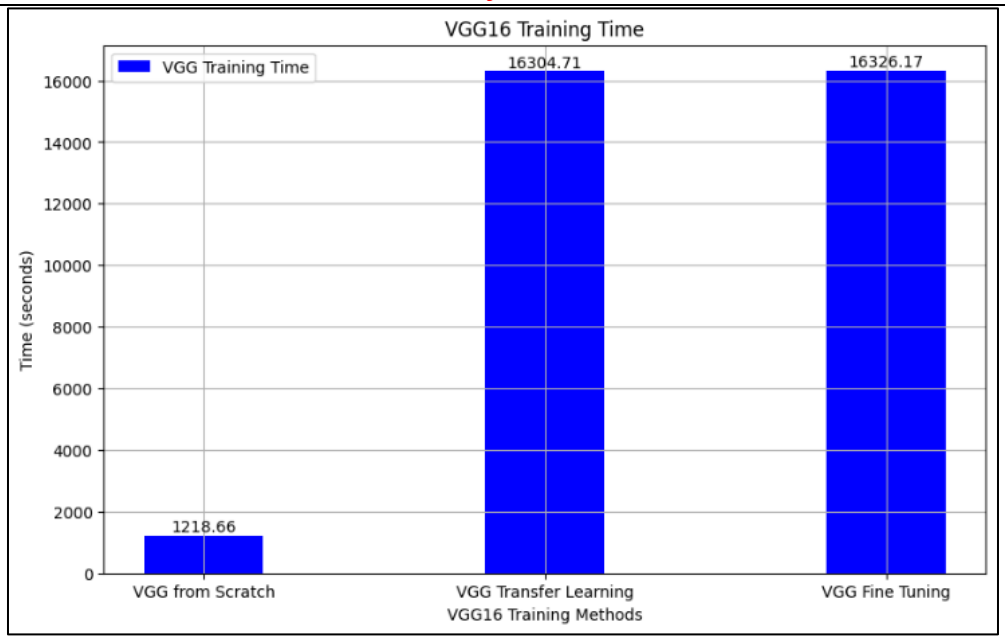
Hence, the result of the analysis is to conclude that VGG16 is more effective than ResNet50 for most of the training techniques, especially transfer learning. This proves that because of the use of pre-trained weights, VGG16 outperforms while ResNet50 appears to have some issues when it comes to transfer learning. Furthermore, both architectures show that fine-tuning can lead to a decline in performance; therefore, training methodologies should be selected carefully. Thus, these results suggest that there is a need to choose the right method for model training that will maximize the results given the data and the problem at hand.

In Figure 22 when presenting these results with bar graphs, it is easy to see that one of the architectures is far outperforming the other. In all trials, the results showed that VGG16 was more suitable to adapt to the FER-2013 dataset especially when the images were in grayscale than ResNet50.

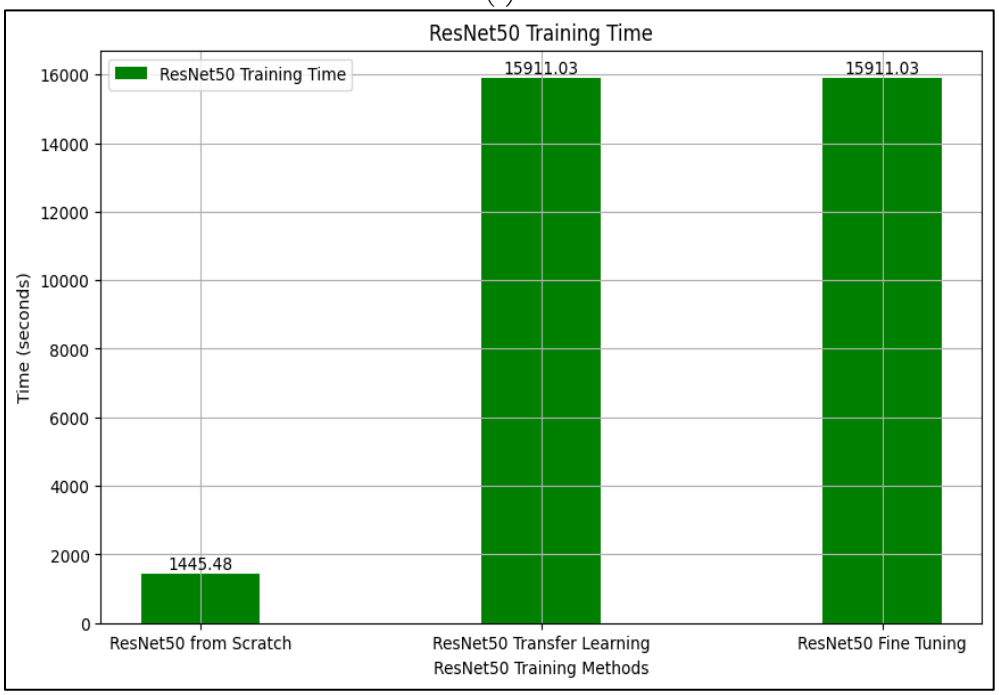
Training Time Comparison:

Training times of VGG16 and ResNet50 are different from each other in terms of various methods. The training time of VGG16 from scratch is 1092.51 sec, transfer learning is 6651.64 sec, and fine-tuning is 12329.30 sec. ResNet50 takes 1,445.48s in scratch training and around 15,911.03s in both transfer learning and fine-tuning as presented in Figure 23.

This implies that ResNet50 needs more computations especially when fine-tuning than VGG16 because of its deeper architecture. Even though the VGG16 takes a shorter time to train both from scratch and when fine-tuning, ResNet50 takes a longer time to train because of its high complexity and computational intensity for deeper models.



(a)



(b)

Figure 23. Comparison of **a.** VGG16 Training Time **b.** ResNet50 Training Time

Table 5. Training Times for VGG16 and ResNet50 Models (in seconds)

Training Approach	VGG16	Res Net 50
From scratch	1218.66	1445.48
Transfer learning	16304.71	15811.03
Fine-tuning	16326.17	15811.03

In Table 5 all the approaches, the training rates of VGG16 are higher than those of ResNet50. The results indicate that both model architecture and training method significantly influence the amount of time required for the computation. In transfer learning and fine-tuning cases, ResNet50 takes more than twice the time VGG16 takes to train. Since using scratch training and moving to transfer learning or fine-tuning increases the training time of both

models, these strategies, although beneficial in improving the model's performance, are computationally costly. Specifically, the applicability of training strategy in resource-constrained settings is a function of the amount of time that is taken to train and the level of performance improvement that is achieved. It is quite surprising that the training time of both transfer learning and fine-tuning of ResNet50 is the same, and this requires a deeper understanding. This may be due to implementation details, measurement errors, or characteristics of ResNet50 in our context of the dataset and the task.

Table 6. Comparison of Training and Validation Accuracy with Training Time Across Different Models and Methods

Model	Training Method	Training Accuracy	Validation Accuracy	Training Time (seconds)
VGG16	From Scratch	0.7259	0.6723	8764.24
	Transfer Learning	0. 7969	0. 6715	16304.71
	Fine Tuning	0. 4650	0. 4813	16326.17
ResNet50	From Scratch	0.5618	0.5469	5771.00
	Transfer Learning	0.2655	0.2846	5803.00
	Fine Tuning	0.2514	0.2475	5793.00

Table 6 summarizes the Comparison of Training and Validation Accuracy with Training Time across different models and methods.

Conclusion and Future Work:

In Conclusion, the proposed models VGG16 and ResNet50 for facial emotion recognition using the FER-2013 dataset explore training from scratch, transfer learning, and fine-tuning strategies. When trained from scratch, VGG16 achieved a validation accuracy of 67.23%, surpassing benchmarks like ResNet50 and VGGN which achieved 54.67%. FER-2013 has a complex dataset, but simpler architectures can navigate it effectively. By enhancing previous models, our research not only sets a new standard for val. Accuracy, but also overlays the way for future developments. There were limitations to our study due to the FER-2013 dataset's challenges, including image quality variations and mixed content (cartoons and human images), which impacted model performance. We were also unable to optimize the Reduce LROn Plateau callback exhaustively due to limited computational resources.

In future research, it would be beneficial to explore a greater number of datasets and improved hyperparameter optimization techniques. To make models more accurate and robust, novel architectures, data augmentation, and preprocessing strategies could also be explored.

Acknowledgment: On behalf of the research team, the authors would like to express the deepest gratitude to all the academic institutions, mentors, and peers who provided their invaluable input during this research.

Author's Contribution: All authors contributed equally to the conceptualization, design, implementation, and writing of this manuscript.

Conflict of Interest: The authors declare no conflict of interest regarding the publication of this manuscript in IJIST.

References:

[1] M. Bansal, M. Kumar, M. Sachdeva, and A. Mittal, "Transfer learning for image classification using VGG19: Caltech-101 image data set," J. Ambient Intell. Humaniz. Comput., vol. 14, no. 4, pp. 3609–3620, Apr. 2023, doi: 10.1007/S12652-021-03488-Z/TABLES/8.

[2] A. Voulodimos, N. Doulamis, A. Doulamis, and E. Protopapadakis, "Deep Learning for Computer Vision: A Brief Review," Comput. Intell. Neurosci., vol. 2018, 2018, doi: 10.1155/2018/7068349.

[3] M. R. Appasaheb Borgalli and D. S. Surve, "Deep learning for facial emotion recognition

using custom CNN architecture," J. Phys. Conf. Ser., vol. 2236, no. 1, p. 012004, Mar. 2022, doi: 10.1088/1742-6596/2236/1/012004.

[4] "Deep Learning." Accessed: Oct. 23, 2024. [Online]. Available: https://www.deeplearningbook.org/

[5] S. Yousefi, M. P. Nguyen, N. Kehtarnavaz, and Y. Cao, "Facial expression recognition based on diffeomorphic matching," Proc. - Int. Conf. Image Process. ICIP, pp. 4549–4552, 2010, doi: 10.1109/ICIP.2010.5650670.

[6] H. Huang, "A facial expression recognition method based on Convolutional Neural Network," Front. Comput. Intell. Syst., vol. 2, no. 1, pp. 116–119, Nov. 2022, doi: 10.54097/FCIS.V2I1.3178.

[7] K. Sarvakar, R. Senkamalavalli, S. Raghavendra, J. Santosh Kumar, R. Manjunath, and S. Jaiswal, "Facial emotion recognition using convolutional neural networks," Mater. Today Proc., vol. 80, pp. 3560–3564, Jan. 2023, doi: 10.1016/J.MATPR.2021.07.297.

[8] L. Zahara, P. Musa, E. Prasetyo Wibowo, I. Karim, and S. Bahri Musa, "The Facial Emotion Recognition (FER-2013) Dataset for Prediction System of Micro-Expressions Face Using the Convolutional Neural Network (CNN) Algorithm based Raspberry Pi," 2020 5th Int. Conf. Informatics Comput. ICIC 2020, Nov. 2020, doi: 10.1109/ICIC50835.2020.9288560.

[9] J. Xiang and G. Zhu, "Joint face detection and facial expression recognition with MTCNN," Proc. - 2017 4th Int. Conf. Inf. Sci. Control Eng. ICISCE 2017, pp. 424–427, Nov. 2017, doi: 10.1109/ICISCE.2017.95.

[10] "Facial Emotion Detection Using Neural Network." Accessed: Oct. 23, 2024. [Online]. Available: https://www.researchgate.net/publication/344331972_Facial_Emotion_Detection_Using_Neural_Network

[11] M. Mukhopadhyay, S. Pal, A. Nayyar, P. K. D. Pramanik, N. Dasgupta, and P. Choudhury, "Facial Emotion Detection to Assess Learner's State of Mind in an Online Learning System," ACM Int. Conf. Proceeding Ser., pp. 107–115, Feb. 2020, doi: 10.1145/3385209.3385231.

[12] D. K. R. Gaddam, M. D. Ansari, S. Vuppala, V. K. Gunjan, and M. M. Sati, "Human Facial Emotion Detection Using Deep Learning," Lect. Notes Electr. Eng., vol. 783, pp. 1417–1427, 2022, doi: 10.1007/978-981-16-3690-5_136.

[13] S. Singh and F. Nasoz, "Facial Expression Recognition with Convolutional Neural Networks," 2020 10th Annu. Comput. Commun. Work. Conf. CCWC 2020, pp. 324–328, Jan. 2020, doi: 10.1109/CCWC47524.2020.9031283.

[14] R. M. D, H. H. Kenchannavar, and U. P. Kulkarni, "Facial Emotion Recognition using Three-Layer ConvNet with Diversity in Data and Minimum Epochs," Int. J. Intell. Syst. Appl. Eng., vol. 10, no. 4, pp. 264–268, Dec. 2022, Accessed: Oct. 23, 2024. [Online]. Available: https://ijisae.org/index.php/IJISAE/article/view/2225

[15] M. Coşkun Irmak, M. Bilge Han Taş, S. Turan, and A. Haşıloğlu, "Emotion Analysis from Facial Expressions Using Convolutional Neural Networks," Proc. - 6th Int. Conf. Comput. Sci. Eng. UBMK 2021, pp. 570–574, 2021, doi: 10.1109/UBMK52708.2021.9558917.

[16] D. Bhagat, A. Vakil, R. K. Gupta, and A. Kumar, "Facial Emotion Recognition (FER) using Convolutional Neural Network (CNN)," Procedia Comput. Sci., vol. 235, pp. 2079–2089, Jan. 2024, doi: 10.1016/J.PROCS.2024.04.197.

[17] K. Lalli and M. Senbagavalli, "Enhancing Deep Learning for Autism Spectrum Disorder Detection with Dual-Encoder GAN-based Augmentation of Electroencephalogram Data," Salud, Cienc. y Tecnol. - Ser. Conf., vol. 3, pp. 958–958, Jan. 2024, doi: 10.56294/SCTCONF2024958.

[18] "FER-2013." Accessed: Oct. 23, 2024. [Online]. Available: https://www.kaggle.com/datasets/msambare/fer2013

[19] H. D. Nguyen, S. Yeom, G. S. Lee, H. J. Yang, I. S. Na, and S. H. Kim, "Facial Emotion Recognition Using an Ensemble of Multi-Level Convolutional Neural Networks," Int. J. Pattern Recognit. Artif. Intell., vol. 33, no. 11, Oct. 2019, doi: 10.1142/S0218001419400159.

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

[21] "Exploring ResNet50: An In-Depth Look at the Model Architecture and Code Implementation | by Nitish Kundu | Medium." Accessed: Oct. 12, 2024. [Online]. Available: https://medium.com/@nitishkundu1993/exploring-resnet50-an-in-depth-look-at-the-model-architecture-and-code-implementation-d8d8fa67e46f