Comparative Analysis of Feature Engineering and Deep Learning Techniques for Robust Facial Emotion Recognition in Real-world Conditions

Claire Russell

Abstract. Facial Emotion Recognition (FER) is valuable in many areas, such as healthcare and human-computer interaction. This study explores FER by comparing traditional classifiers, such as Support Vector Machine (SVM), with current deep learning methods like Convolutional Neural Networks (CNNs). Employing the Static Facial Expressions in the Wild (SFEW) dataset, the investigation underscores the challenges presented by real-world scenarios. While SVMs set a foundational benchmark, the CNNs, especially when fine-tuned with VGG-16, showcased a significant leap in performance, achieving an accuracy of 44.12%. Despite these advancements, the journey towards achieving robust, real-world applicability in FER remains ongoing, with future research avenues highlighted.

Keywords: Facial Emotion Recognition (FER), Support Vector Machine (SVM), Convolutional Neural Networks (CNN), Static Facial Expressions in the Wild (SFEW), Deep learning, VGG-16 fine-tuning, Real-world Conditions, Image preprocessing, Feature engineering, Model optimisation, Model Tuning.

1 Introduction

Facial emotion recognition (FER) is crucial in many fields such as human-computer interaction and healthcare. The goal is to accurately interpret facial expressions into emotions. This can help computers and devices respond better to human needs. For instance, in healthcare, detecting a patient's stress levels can guide their treatment, especially in cases where they may be nonverbal. Similarly, in education, understanding a child's emotional response to an activity can help tailor the learning experience, ensuring challenges without overwhelming them. The Static Facial Expressions in the Wild (SFEW) dataset was chosen for this investigation due to its reflection of real-world challenges such as diverse head poses, occlusions, and uncontrolled illumination conditions, which are known to affect the performance of FER systems.

The task at hand is to classify facial expressions into one of seven distinct emotions. This study will compare the effectiveness of two distinct datasets and classification techniques. The first dataset has been derived through feature engineering, while the second consists of the raw images themselves. Initially, a Support Vector Machine (SVM) model, as referenced in paper [1], is utilised for benchmarking. For many years, SVMs were the preferred classical classification method for FER. However, with advancements in deep learning, Convolutional Neural Networks (CNNs) have emerged as the leading approach, as highlighted in reference [3]. After the baseline CNN is implemented, a pretrained model will be fine-tuned to see if it offers better performance. Fine-tuning a pretrained model can leverage already learnt features from vast datasets, potentially improving accuracy by building upon prior knowledge instead of starting from scratch. The goal is to determine which combination of dataset and classification method, including the fine-tuning approach, proves most effective for facial emotion recognition in real-world scenarios.

2 Method

2.1 Data inspection

After cleaning the feature-engineered dataset by removing a row with NaN values, the dataset now comprises 674 samples, sourced from 38 distinct movies. It has seven class labels: angry, disgust, fear, happy, neutral, sad, and surprised. The class distribution is mostly balanced with 100 samples in each, except for fear with 99 and disgust with 75. Each sample is represented by a ten-dimensional feature vector, comprising the first five principal components of Local Phase Quantization (LPQ) and Pyramid of Histogram of Gradients (PHOG) features, which have been effective in lab-controlled FER datasets. LPQ is a robust descriptor for capturing textures in face images even with blurring, which is key for an “in the wild” dataset whereas PHOG features provide information on shape and spatial layout [2]. Despite the initial similarity in feature scales, they were standardised to ensure consistency and numerical stability during training.

The class-wise histograms for each standardised feature show variations in the distribution of values across different emotions, suggesting these features could be useful for emotion classification. For example, a notable peak in PHOG_2 for "Disgust" and a right-skewed distribution in PHOG_3 for "Fear" may indicate their discriminative power for these emotions. However, overlapping distributions in multiple features across emotions hint at potential challenges in distinguishing between different emotions. This raises a question on whether more distinct class-wise patterns would

emerge in lab-controlled FER datasets.

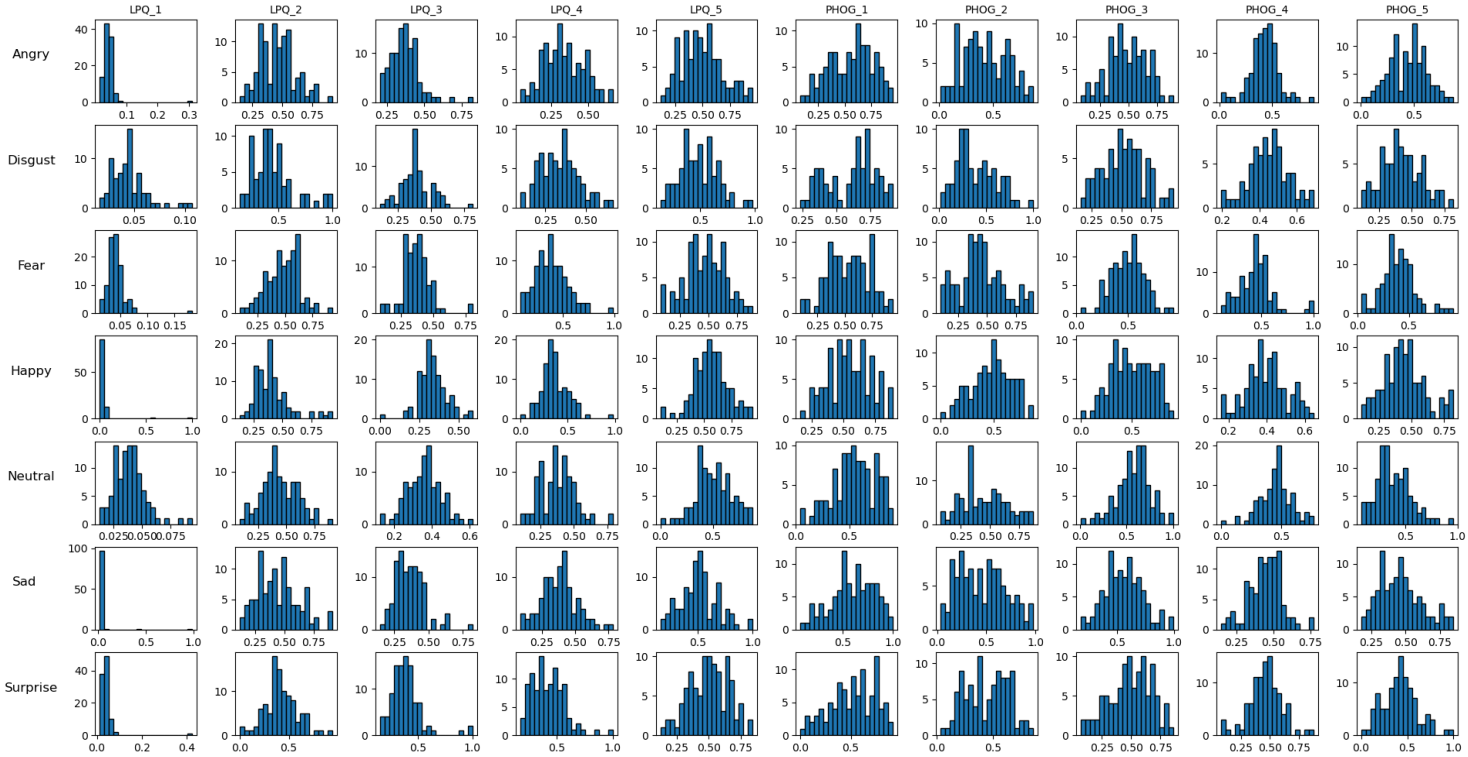

Fig. 1. Fig. 1. Class-wise histograms depicting the distribution of feature values for each emotion. Each row corresponds to a distinct emotion, while each column represents a different feature.

2.2 SVM Classification

A baseline for classification accuracy was established by replicating the method from the referenced paper, employing a Support Vector Machine (SVM) with a Radial Basis Function (RBF) kernel and five-fold cross-validation as outlined in script [1]. A high regularization parameter C=1000 and scale gamma were selected using grid search. However, such a high C value could potentially lead to model overfitting on the training data.

Similar challenges to the original authors were encountered in achieving high accuracy on the dataset, with an average accuracy of 23.89% observed over the five folds. Emotions such as Disgust, Surprise, and Neutral yielded the three lowest precision and recall scores, while Angry, Fear, and Happy were somewhat easier for the SVM to classify. This variance in classification performance across emotions could be attributed to inherent complexities in distinguishing certain emotions, varying effectiveness of the features in capturing the characteristics of different emotions, and overlap between classes in the feature space (e.g., Angry and Surprise). The class-wise confusion matrix, consolidated over the five folds, is depicted below.

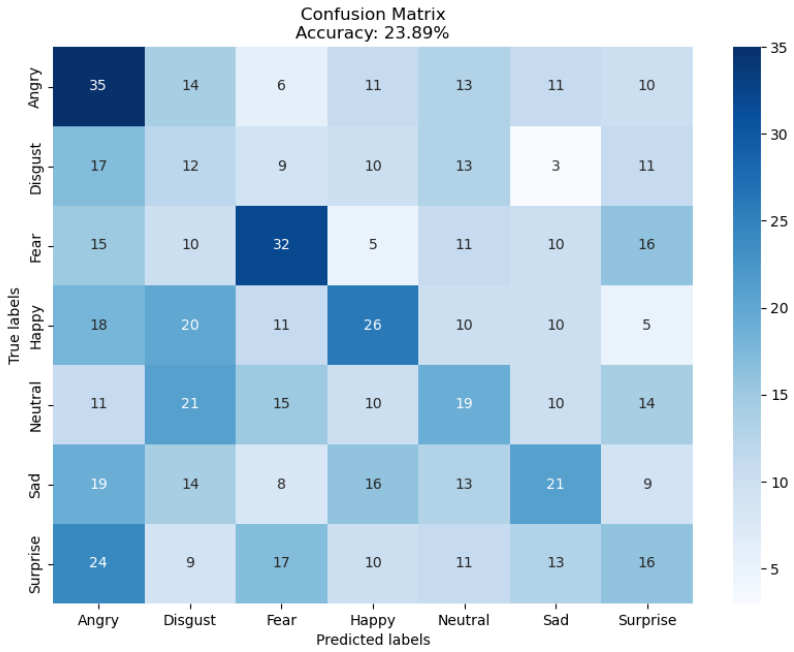

Figure 2. (Previous page) Class-wise Confusion Matrix illustrating the performance of SVM in Facial Emotion Recognition (FER) classification across various emotions.

Table 1. Class-wise Precision and Recall Metrics for SVM-based Facial Emotion Recognition (FER) Classification.

Emotion	Precision	Recall
Angry	25.18%	35.00%
Disgust	12.00%	16.00%
Fear	32.65%	32.32%
Happy	29.55%	26.00%
Neutral	21.11%	19.00%
Sad	26.92%	21.00%
Surprise	19.75%	16.00%

2.3 Neural Net classifier

After setting a baseline with an SVM classifier, a neural network was used to try and improve the results. A simple feedforward neural network with two hidden layers, each having 64 neurons, was set up. The ReLU activation function was used in the hidden layers. The output layer provided raw class scores (logits) without applying a softmax activation, as the later-used loss function (nn.CrossEntropyLoss) already includes softmax processing.

A grid search was conducted over specified ranges for learning rate, hidden layer sizes, number of epochs, and batch size to find the optimal configuration on a validation split. Although easy to implement, because grid search operates on a discrete set of values for each hyperparameter, it may potentially overlook optimal values that fall between the specified grid points. Experiments were also carried out with different activation functions and the inclusion of a dropout layer to reduce overfitting. The model was trained for 1000 epochs without mini-batches (i.e., full batch training using all 539 training examples) and learning rate of 0.01.

Full batch training provides stable and clear convergence towards the minimum of the loss function since the entire dataset is used to compute the gradient at each iteration. However, it can be computationally expensive and slow, especially for large datasets or complex models. In this case the dataset was relatively small and full batch training did not drastically increase run time.

Despite some configurations reaching lower loss values during training, the validation accuracies were not particularly high, with the highest being around 35.19%. This pattern suggests that the model may be overfitting to the training data, learning to memorise the training examples rather than generalising from them to perform well on unseen data.

Stratified 5-fold cross-validation was utilised to ensure a more reliable assessment of the neural network's performance across various subsets of the dataset, thereby enhancing the evaluation's robustness. This approach guarantees that every data point is used for both training and testing, which is crucial in the context of a limited dataset. Moreover, at each split of the data, a balanced representation of classes was maintained, ensuring that the model learns to recognise each emotion effectively, without bias towards over-represented classes, leading to a more accurate and reliable evaluation.

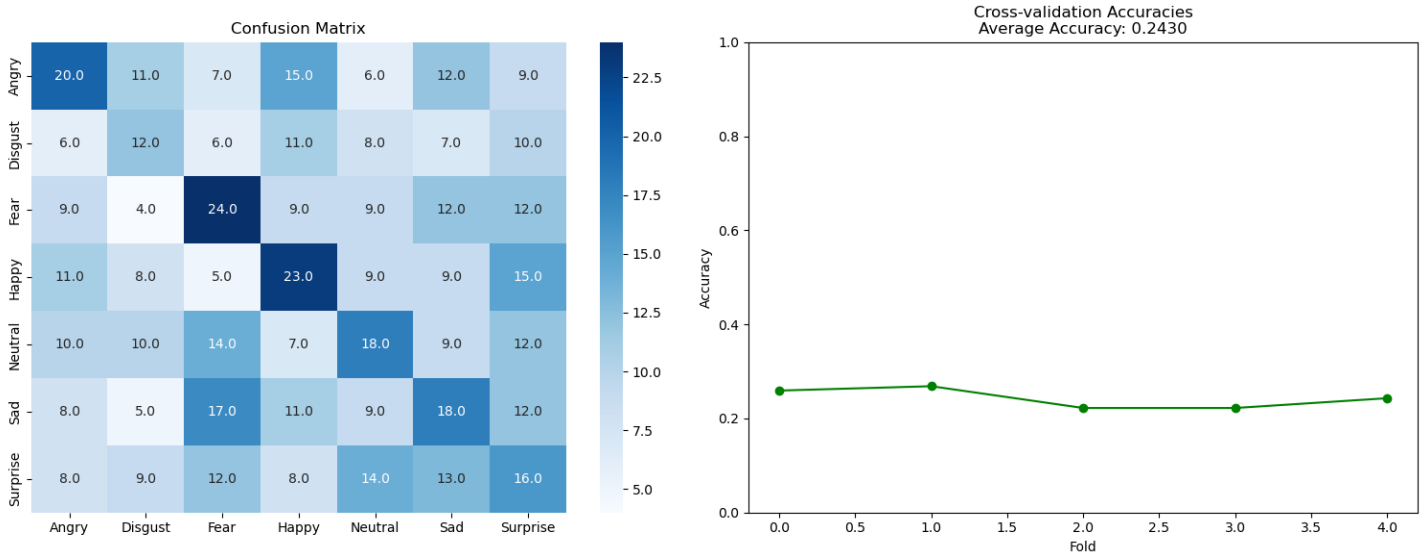

Figure 3. Class-wise Confusion Matrix depicting the performance of a Neural Network in Facial Emotion Recognition (FER) classification across various emotions, accompanied by the 5-fold Cross-Validation accuracies.

In the cross-validation stage, the neural network's performance varied slightly across different folds with a mean accuracy of 24.30%, approximately equal to the SVM. Examining the class breakdown, it appears that the Neural Network demonstrates better consistency across emotions, though this result would need to be validated through further replication of the experiment. This consistency could be attributed to the greater complexity of Neural Networks compared to SVMs, allowing them to learn more complex patterns in the data. Alternatively, it could be related to the random initialisations and stochastic nature of neural networks, which can lead to variations in performance across different runs. The line graph demonstrates that the model's performance was consistent across different subsets of the data.

Table 2. Class-wise Precision and Recall Metrics for Neural Network-based Facial Emotion Recognition (FER) Classification.

Emotion	Precision	Recall
Angry	24.71%	25.00%
Disgust	22.90%	20.00%
Fear	28.72%	30.42%
Happy	28.03%	28.75%
Neutral	23.11%	22.50%
Sad	23.80%	22.50%
Surprise	19.41%	20.00%

The model was then trained on the entire training dataset using the best hyperparameters with 20 % of the data reserved for testing. The data used for testing was excluded from hyperparameter tuning and cross validation to ensure it was "unseen" by the model. Finally, the model was evaluated on a held-out test set to obtain a final estimate of its performance. The final evaluation on the test data yields an accuracy of 26.67%, which again is low. Although the model displayed quite a lot of instability over various runs of the program, Happy was consistently the easiest emotion for both the SVM and Neural Network to classify.

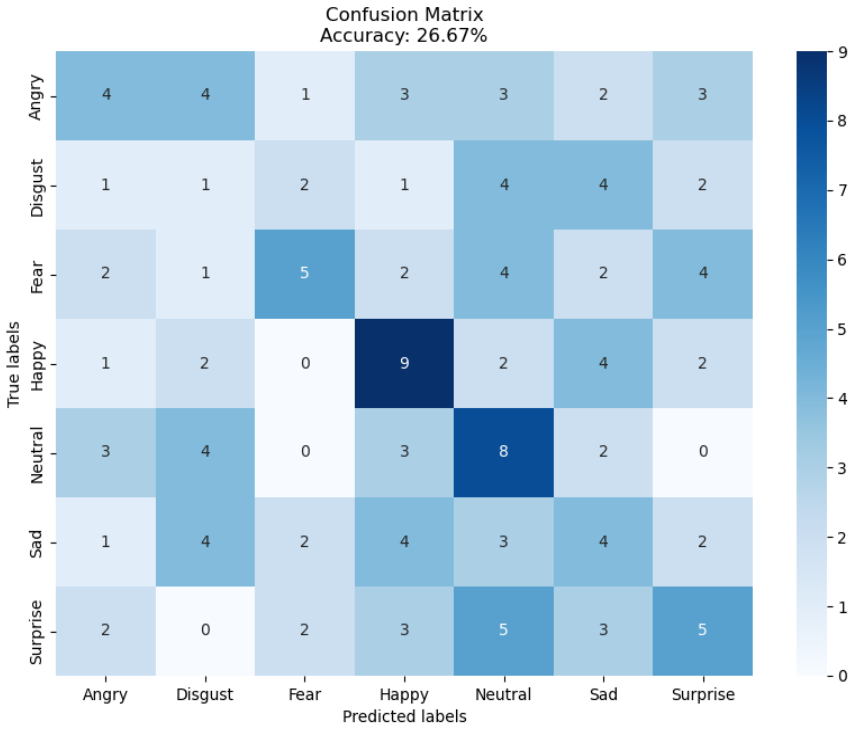

Figure 4. Class-wise Confusion Matrix illustrating the performance of a Neural Network in Facial Emotion Recognition (FER) classification across various emotions over the test set.

2.4 Image Preprocessing for CNN-based Emotion Recognition

A critical step in preparing data for facial emotion recognition (FER) using convolutional neural networks (CNN) is isolating the face from each image. For this purpose, the dataset's images were cropped to highlight the face. As in the featured engineered dataset, there were 100 samples for all emotions except Disgust, for which there were 75. The 'mtcnn' Python package, which utilises Multi-task Cascaded Convolutional Networks for face detection based on TensorFlow, was primarily employed. However, faces in approximately 25 images were challenging to detect due to factors like lighting variations, unusual angles, or partial occlusions of the face by hair, hands etc. These were manually cropped. This cropping process is vital as it directs the model's attention to relevant facial features, minimising

distractions from the background. In addition to providing consistent input to the CNN, this also reduces computational demands by focusing on smaller, face-centric image sections.

Post-cropping, the images undergo on-the-fly transformations during training. This means that every time an image is passed through the network during different epochs, it might be subjected to slight alterations due to transformations like random horizontal flipping and random rotation. Such preprocessing techniques are routinely employed in FER tasks [4]. Implementing on-the-fly transformations ensures that the model sees varied versions of the same image, hopefully enhancing its generalisation capabilities. Experiments with other dimensions, such as 96×96 and 128×128 pixels, were conducted in both colour and greyscale. However, the results indicated superior performance with the larger, colour images. While it might initially seem that colour is not crucial for FER, and the training samples below do not immediately suggest subtle cues of cues like blush or pallor were present in the data set, having three channels instead of one appears to convey some additional information.

While several additional augmentation techniques were experimented with, such as random cropping and random greyscale conversion, there was a noticeable plateau in performance improvement after a certain point. This suggests that while augmentations can increase the robustness of the model, over-augmenting can lead to the model focusing on less relevant features or, in some cases, might even introduce noise. It is a delicate balance to strike: while diversity in training data is advantageous, it is essential to ensure that this diversity remains representative of real-world scenarios and does not deviate from the primary task of emotion recognition.

Figure 5. 10 samples from the training dataset represent each of the seven classes after cropping and data augmentation. Some images may appear flipped or subtly rotated, while others exhibit variations in colour saturation and tone due to the back transformation of original RGB values following normalisation.

2.4 Optimising CNN Architecture and Hyperparameters for Classification

Optimising a Convolutional Neural Network (CNN) for facial emotion recognition (FER) requires careful consideration of both the model architecture and hyperparameters. Given the limited size of the dataset, the overarching challenge was ensuring the model could adequately learn from the data without overfitting. This balancing act between model regularisation and its capacity to detect features was crucial.

In the final approach, three-layer CNN was employed, with batch normalisation applied after each convolution to stabilise learning and increase training speed. The network utilised max-pooling to progressively reduce the spatial dimensions of the data, allowing for higher-level feature detection in the deeper layers. However, the true balancing act came in the form of hyperparameter tuning.

Learning rate, weight decay, and batch size are some of the most influential hyperparameters in a neural network's training process. A smaller learning rate of 1×10^{-5} was chosen, enabling the model to make incremental adjustments and

hopefully find a more optimal solution. Weight decay, a regularisation technique, was introduced to penalise larger weights in the model, thereby discouraging overfitting. Its value was set at 1×10^{-4}. Although dropout is a common regularisation technique, introducing it at various layers from vales ranging from 5% - 50% did not enhance performance. This could be due to the already regularising effect of the batch normalisation layers.

Kernel sizes and strides were extensively evaluated. Some literature suggested that larger kernel sizes (up to 8) might be more effective [4]. However, after assessing various configurations, a kernel size of 3 and a stride of 2 were chosen based on their performance with the dataset. Various epochs, early stopping criteria, and different batch sizes were also experimented with. Early stopping can be a powerful tool, halting training when the model starts to overfit to the training data. However, even with these techniques in play, extensive hyperparameter tuning could not increase the accuracy on the test set beyond 40%. Similar accuracy was achieved on different splits of the training and validation data. While this accuracy is respectable for a seven-class classification problem, especially given the challenges of the dataset, it is not ideal. It is possible that training the final model on both the training and validation sets combined could have yielded better results, but time constraints did not allow for this test.

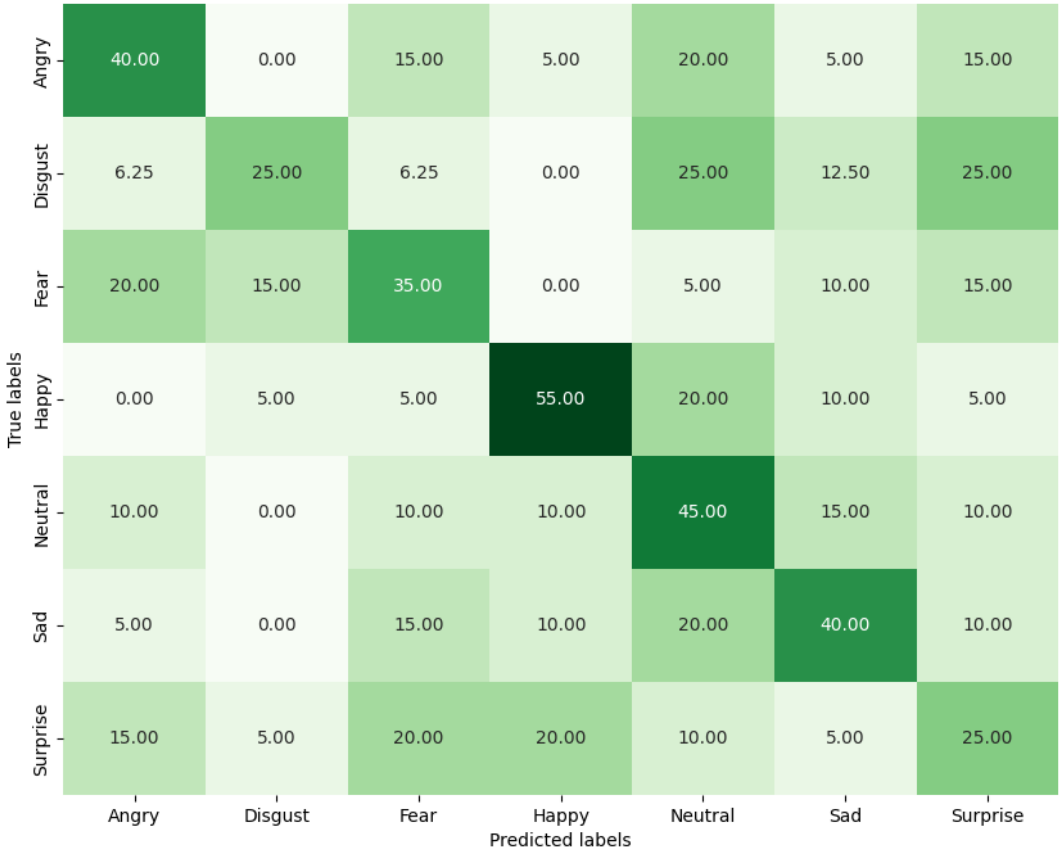

Figure 6. Class-wise Normalised Confusion Matrix showcasing the CNN's performance in Facial Emotion Recognition (FER) on the test set. Each cell reflects the percentage of true labels within each emotion category, with row values totalling 100%. Like the SVM and NN evaluations, the emotion 'disgust' proved challenging to classify. It's important to note these results come from a single run. While some trends (like 'happy' consistently performing well) remained, there was notable variation between runs, suggesting some instability.

```
Epoch 47/50
Train Loss: 0.3655, Train Accuracy: 92.58%
Validation Loss: 1.8592, Validation Accuracy: 44.78%
------------------------------
Epoch 48/50
Train Loss: 0.3600, Train Accuracy: 94.92%
Validation Loss: 1.8263, Validation Accuracy: 41.79%
------------------------------
Epoch 49/50
Train Loss: 0.3407, Train Accuracy: 95.13%
Validation Loss: 1.8565, Validation Accuracy: 41.79%
------------------------------
Epoch 50/50
Train Loss: 0.3301, Train Accuracy: 95.13%
Validation Loss: 1.8715, Validation Accuracy: 43.28%
------------------------------
Test Accuracy: 38.24%
```

Figure 7. Output from the final few training epochs indicating overfitting: While training accuracy continues to increase, validation accuracy plateaued around the 20th-30th epoch. Accuracy on the test set for this run was 38.24%.

2.5 Model Fine-Tuning Using VGG-16

Finally, a pre-trained VGG-16 model was adapted to see if results could be improved. The VGG-16 model, a renowned deep convolutional neural network, has shown exemplary performance in large-scale image recognition tasks. It was trained on a large-scale dataset that consists of over 14 million images spanning 1,000 different classes, including those related to everyday objects, animals, and scenes. The dataset has been instrumental in advancing the field of deep learning, particularly in the context of image classification tasks. Leveraging a pre-trained model presents considerable advantages, especially with limited data. This approach harnesses the model's capacity to build upon previously learned features from expansive datasets like ImageNet. Emotion recognition using fine tuned models is well-established method in the literature [5].

The VGG-16 model was chosen as a starting point due to its proven performance in image tasks. Initially, all its layers were frozen to ensure that the inherent features, learned from vast datasets, were retained. The model was then adapted to address the specific requirements of the FER task. This involved altering the last fully connected layer to correspond with the seven emotion classes found in the FER dataset. Once this adjustment was made, the fully connected layers, commonly known as the classifier section of VGG-16, were set to be trainable, paving the way for fine-tuning on the FER dataset.

The training process was carried out over several epochs. At each stage, the model's performance metrics, including loss and accuracy, were closely monitored on both the training and validation datasets. This iterative evaluation provided insights into the model's learning progression. Upon completing the training, the model was put to the test, and it achieved an accuracy of 44.12%. While this initial outcome shows potential, it is essential to run more tests to confirm the model's consistent and reliable performance.

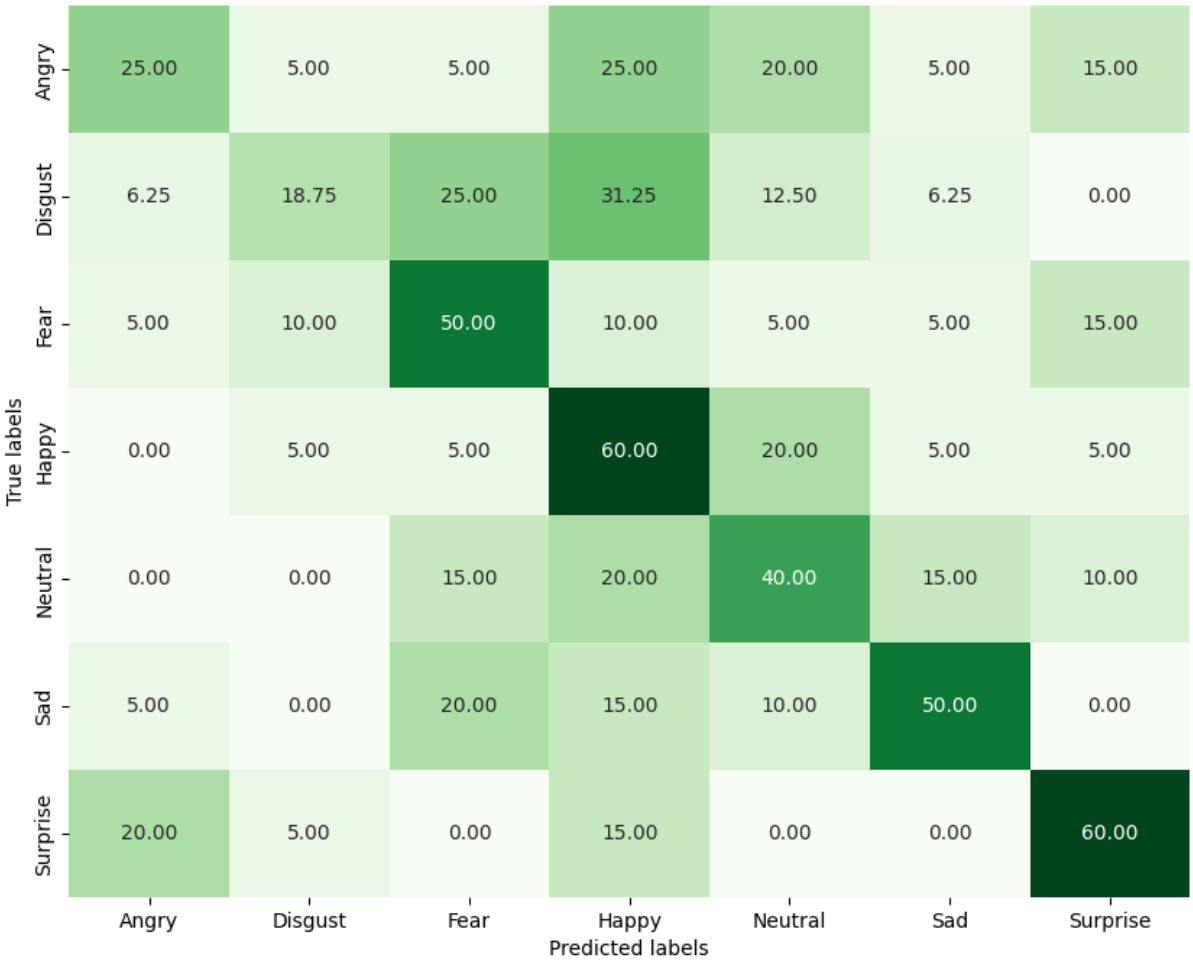

Figure 7. Normalised Confusion Matrix for the VGG-16 fine-tuned CNN model, illustrating performance in Facial Emotion Recognition (FER) on the test set

3 Results and Discussion

Various classifiers were explored for facial emotion recognition (FER), including the traditional Support Vector Machine (SVM), basic neural networks, and the Convolutional Neural Network (CNN). Initial experiments with SVM and the simple neural network produced modest results, with accuracies roughly in the mid-twenties. In contrast, the CNN, even prior to intensive optimisation, demonstrated a marked improvement, nearing 40% accuracy on the test set.

A significant leap in performance was observed when the CNN architecture was further fine-tuned using the VGG-16 model. This adaptation allowed the model to achieve an accuracy of 44.12% on the test set. The VGG-16 model, being pretrained on the expansive ImageNet dataset, brought with it a wealth of learned image features. By fine-tuning it for the FER task, the model could leverage these generic features and hone them for emotion recognition.

The inherent strength of CNNs lies in their ability to automatically learn hierarchical features from images, negating the need for manual feature extraction. This characteristic allows CNNs to discern intricate facial cues and patterns that might prove too difficult for traditional methods. Specifically, the convolution layers in CNNs excel at identifying both basic features like edges and textures, as well as abstract, high-level features pivotal for emotion detection.

Despite the advancements achieved with CNN and fine-tuning, the results, while promising, still fall short for practical, real-world applications. The "in the wild" nature of the dataset, with its myriad of challenges like variable lighting, occlusions, and diverse facial orientations, greatly increases the complexity of the task. The limited size and multifaceted nature of the dataset might have set a ceiling on the models' capabilities, restricting them from reaching optimal performance.

Looking forward, there is substantial scope for enhancement. Delving into advanced regularisation techniques, exploring ensemble methods, or harnessing the potential of semi-supervised learning with unlabelled data could pave the way for significant performance improvements. Additionally, integrating facial landmarks or other facial analysis techniques might offer complementary cues, bolstering the model's emotion recognition ability.

4 Conclusion and Future Work

This study provided a comprehensive analysis of facial emotion recognition (FER) through the comparison of classical and deep learning classifiers. While traditional methods such as the Support Vector Machine (SVM) set a foundational benchmark, the advancements in deep learning, especially Convolutional Neural Networks (CNN), showcased their superiority in handling such complex tasks. The fine-tuning of the VGG-16 model further demonstrated the potential benefits of leveraging pre-trained architectures, as it led to a notable improvement in FER performance.

One promising avenue for further enhancement is the integration of evolutionary algorithms in optimising the architecture of CNNs. Evolutionary algorithms, inspired by the process of natural selection, can be adept at navigating vast search spaces to find optimal or near-optimal solutions. In the context of CNNs, they could be harnessed to fine-tune architectural elements such as the number of layers, the number of neurons in each layer, kernel sizes, and activation functions, among others. By evolving the architecture iteratively, the model can potentially adapt better to the nuances of the dataset, leading to improved performance. Given the sensitively of the model seen during training to hyperparameter settings, this would be a useful avenue for future research in this space.

Besides architectural optimisation, incorporating other modalities like audio cues or physiological signals could provide complementary information, enhancing the robustness of emotion recognition systems. Additionally, integrating advanced feature extraction techniques, such as facial landmarks or depth maps, might yield richer data representations, further boosting recognition accuracy.

In the future, it would be beneficial to explore larger and more diverse datasets, possibly harnessing the power of transfer learning from multiple pre-trained models. Combining traditional feature-based approaches with deep learning could also provide a more holistic view, capitalising on the strengths of both methodologies.

In conclusion, while significant strides have been made in the domain of facial emotion recognition, the journey towards achieving robust, real-world applicability is ongoing. The insights gained from this study pave the way for future research, with the hope that continuous innovations will lead to more empathetic and intuitive human-computer interactions.

References

1. Dhall, A., Goecke, R., Lucey, S., & Gedeon, T.: Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark. In: 2011 IEEE international conference on computer vision workshops (ICCV workshops), pp. 2106-2112. IEEE, (2011).
2. Dhall, A., Asthana, A., Goecke, R., Gedeon, T.: Emotion recognition using PHOG and LPQ features. In: 2011 IEEE International Conference on Automatic Face & Gesture Recognition (FG), pp. 878-883, IEEE, Santa Barbara, CA, USA, (2011).
3. Canal, F.Z., Müller, T.R., Matias, J.C., Scotton, G.G., de Sa Junior, A.R., Pozzebon, E., Sobieranski, A.C.: A survey on facial emotion recognition techniques: A state-of-the-art literature review. Inf. Sci. 582, 593--617 (2022).
8. Bodapati, J.D., Srilakshmi, U., & Veeranjaneyulu, N.: FERNet: A Deep CNN Architecture for Facial Expression Recognition in the Wild. J. Inst. Eng. India Ser. B 103(3), 439–448 (2022).
5. Kusuma, G.P., Jonathan, J., & Lim, A.P.: Emotion Recognition on FER-2013 Face Images Using Fine-Tuned VGG-16. Adv. Sci. Technol. Eng. Syst. J. 5(2), 315-322 (2020).