Deep Learning Models Comparison for Emotion Classification With Image Pre-Processing Methods

Bryan Anthony[1], Nicholas D. Lienardi[1], Richard E. Sutanto[1], and Yuwono M. Dinata[1]
[1]Informatics Department, Faculty of Science and Technology, Universitas Ciputra Surabaya, Surabaya, Indonesia

Corresponding author: Bryan Anthony (e-mail: bryan.wansen@gmail.com).

ABSTRACT This research investigates advancements in Facial Expression Recognition (FER) within the domain of affective computing, focusing on improving the accuracy and robustness of FER systems under diverse, real-world conditions. Facial expressions serve as critical non-verbal cues in human communication, yet existing FER systems often face challenges due to environmental variability such as changes in lighting, pose, and occlusions. This study evaluates the performance of three Convolutional Neural Network (CNN) architectures—ResNet50, VGG16, and MobileNetV3Large—integrated with preprocessing techniques like Contrast Limited Adaptive Histogram Equalization (CLAHE) and the Synthetic Minority Oversampling Technique (SMOTE). These methods address key challenges such as class imbalance and low contrast in datasets. Results demonstrate the pivotal role of tailored preprocessing strategies. For instance, the application of CLAHE and SMOTE improved the VGG16 model's test accuracy from 0.70 to 0.79, representing a 0.09 or 9% increase. This significant improvement underscores the effectiveness of combining advanced preprocessing methods with CNN architectures. Furthermore, the findings highlight the advantages of optimizing preprocessing to enhance the recognition of subtle emotions in uncontrolled settings, offering practical insights for deploying FER systems in real-time applications. Overall, this research demonstrates the potential of preprocessing techniques to enhance FER system performance significantly, particularly when paired with well-established deep learning models. These insights pave the way for the development of more accurate, robust, and adaptable FER systems capable of functioning reliably in dynamic, real-world environments.

KEYWORDS CLAHE, Computer Vision, FER, SMOTE

I. INTRODUCTION

In psychology, "affect" refers to the outward display of emotions and feelings through non-verbal communication. Non-verbal communication, including facial expressions, eye contact, voice tone and pitch, gestures, and physical distance, is crucial in interpersonal interactions. Research indicates that these non-verbal cues constitute approximately 60% to 80% of communication. Among these, facial expressions are the most commonly analyzed [1]. Affective computing aims to create systems and devices capable of recognizing, interpreting, and mimicking human affects using various channels like facial expressions, voice, and biological signals. Facial expressions, in particular, are a crucial nonverbal way for humans to communicate their internal emotions. Building robust Facial Expression Recognition (FER) systems for Human-Machine Interaction (HMI) has seen significant development, aiming for machines that understand human emotions and respond naturally [2].

However, current Facial Expression Recognition (FER) systems struggle in uncontrolled environments due to factors like lighting, camera angles, and user diversity. To address this issue, achieving a higher level of accuracy is necessary to reliably label these facial expressions. Current research in Facial Affect Recognition (FAR) has explored various methodologies to improve system robustness and accuracy in diverse conditions.

For instance, [3] employed CNNs on large multimedia datasets, achieving high accuracy but facing challenges with real-time processing speeds, which are crucial for practical applications. This highlights the need for further optimization to balance accuracy with processing efficiency. This research [4] integrated attentional mechanisms into CNN models to enhance the recognition of subtle emotions, improving performance. However, their approach may be prone to overfitting, especially when applied to less varied datasets, limiting its generalizability.

This research [5] introduced a hybrid approach that combines traditional feature descriptors with CNNs, which improved feature extraction. Nonetheless, the increased computational complexity may reduce the model's feasibility for real-time deployment. Similarly, [6] developed a hybrid attention cascade network that achieved excellent accuracy in controlled settings, such as the CK+ dataset, but its effectiveness decreases in more variable environments, underscoring the importance of robustness. Finally, [7] focused on driver emotion recognition using CNNs integrated with vehicle sensors. Their work demonstrated innovation in specific contexts but might lack the flexibility needed for broader applications across different environments.

Despite the recent trends of Vision Transformers (ViT) in FER, the challenges associated with training Vision Transformers, particularly in contexts where data is scarce. While ViTs demonstrate promising capabilities in various domains, including Facial Emotion Recognition (FER), they often require extensive datasets and significant computational resources for effective training. This labor-intensive nature can pose barriers, especially in practical applications where quick deployment and efficiency are paramount. As a result, despite the advancements offered by ViTs, traditional Convolutional Neural Networks (CNNs) should still be pursued for FER tasks. Traditional CNNs excel in scenarios with limited data and computational resources, making them more practical for many real-world applications. Their architecture allows for efficient feature extraction from facial images without the need for extensive preprocessing or hyperparameter tuning, which are often necessary for ViTs. CNNs have been widely adopted in FER due to their ability to capture intricate spatial features and their robustness against variations in lighting and facial poses. Moreover, CNNs can be fine-tuned using transfer learning techniques, allowing researchers to leverage pre-trained models on smaller datasets effectively. Thus, while ViTs are an exciting development in the field, traditional CNNs remain essential for achieving reliable and efficient emotion recognition in diverse settings [8].

In the studies using ResNet50, a noted limitation was its high computational cost, especially when applied to large-scale datasets. Although it achieved high accuracy, its deep architecture resulted in longer training times and required substantial hardware resources [9]. VGG16, while effective in feature extraction, faced similar challenges due to its computationally intensive nature, making it less suited for real-time applications [10].

In contrast, MobileNetV3 Large, although more efficient and faster, had slightly lower accuracy compared to deeper architectures like ResNet50. Its lightweight design traded off some depth and complexity, which, while beneficial for mobile devices, reduced its ability to capture very intricate features in FER tasks [11]. Each study highlighted the importance of balancing accuracy and resource efficiency, depending on the specific application requirements.

In summary, while these studies provide valuable insights and advancements in FER, they also reveal ongoing challenges in achieving consistent performance across various real-world conditions. Addressing these issues through improved model generalization, efficiency, and adaptability is crucial for the future development of reliable FER systems.

The contribution of this research is to explore how we could improve further research in Facial Expression Recognition by using different CNN Models in combination with pre-processing methods that can help enrich the state of Facial Expression Recognition in computer vision by creating comparisons between CNN Models performance with several different pre-processing methods.

II. RESEARCH METHODS

A. DATASETS

The experiment utilized the "Facial Expressions Training Data" dataset which was acquired from the Kaggle Website. This dataset consists of 28,175 images in grayscale with a resolution of 96 x 96 pixels categorized into eight classes: surprise, anger, disgust, fear, sadness, contempt, neutral, and happiness. The distribution of images across these classes is as follows: surprise (4,616 images), happiness (4,336 images), anger (3,608 images), disgust (3,472 images), contempt (3,244 images), fear (3,043 images), sadness (2,995 images), and neutral (2,861 images). The distribution of images is illustrated in Figure 1, while example images for each class are provided in Figure 2.

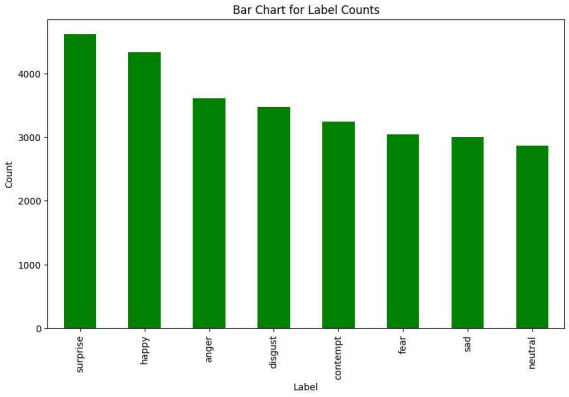

Figure 1. The image distribution on the classes

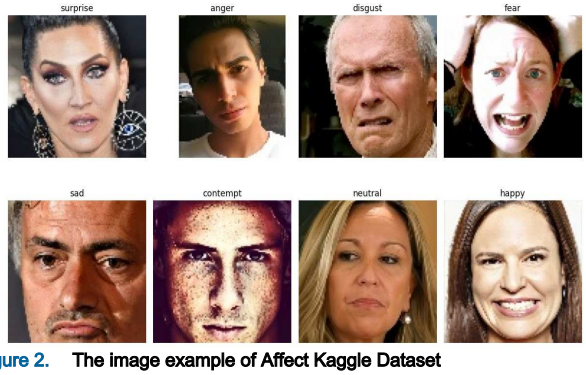

Figure 2. The image example of Affect Kaggle Dataset

This data-set is based on AffectNet-HQ, which used a state-of-the-art FER model to improve on the AffectNet original labels. AffectNet-HQ and its predecessor AffectNet, is a dataset for Facial Expression, Valence, and Arousal Computing in the Wild. The AffectNet dataset was created by researchers from the University of Denver to support facial expression recognition. It consists of over one million images collected via web searches using emotion-related keywords. The dataset is diverse, representing various ages, ethnicities, and genders, making it a valuable resource for developing and evaluating models in affective computing [12].

This dataset served as the primary source for training and testing the proposed models. The data is divided into training, validation, and test sets with a distribution of 80% for training, 10% for validation, and 10% for testing.

B. DATA PREPROCESSING

Data preprocessing is a vital part of data analysis, focused on cleaning, organizing, and readying raw data for further examination. Here's an overview of this stage in the process:

1) DATA CLEANING AND ENHANCEMENT

In the initial stage of preprocessing, Contrast Limited Adaptive Histogram Equalization (CLAHE) was utilized to improve the contrast of grayscale images. This technique enhances the visibility of key facial features, making it easier for models to detect subtle differences in facial expressions. By increasing the contrast in these images, CLAHE ensures that even minor variations in facial features are more distinguishable, which is critical for accurate emotion recognition [13]. SMOTE enhances facial expression recognition (FER) by addressing class imbalance in datasets, which is common in the FERC. By generating synthetic samples for minority classes, SMOTE improves model training on underrepresented facial expressions, leading to better recognition accuracy [14].

2) ADDRESSING CLASS IMBALANCE

To manage the issue of class imbalance within the dataset, the Synthetic Minority Over-sampling Technique (SMOTE) was employed. SMOTE works by generating synthetic samples for underrepresented classes, effectively increasing the number of data points in these classes. This helps in balancing the distribution of facial expressions across the dataset, ensuring that the models do not become biased towards the majority classes, which could otherwise lead to skewed predictions and reduced model accuracy [15]. In their study, [16] highlight the importance of Contrast Limited Adaptive Histogram Equalization (CLAHE) in enhancing facial expression recognition (FER). CLAHE improves image contrast, making subtle facial features more visible, which is crucial for accurate expression classification. It also normalizes lighting conditions across different images, reducing variability that can negatively impact recognition accuracy. By employing CLAHE, models like GoogLeNet-InceptionV3 can achieve better performance in interpreting facial expressions. This preprocessing technique ultimately contributes to the development of more robust and reliable FER systems.

3) UNDERSAMPLING OVERREPRESENTED CLASSES

In addition to oversampling the minority classes, undersampling techniques were applied to reduce the number of samples in overrepresented classes. This step was crucial in preventing the model from being overly influenced by the majority classes. By reducing the dominance of these classes, undersampling contributes to a more equitable training process, allowing the model to learn from a more balanced dataset and thereby improving its ability to generalize across different classes [17].

C. DEEP LEARNING

The experiment involved training three different deep learning architectures to classify facial expressions from the preprocessed dataset. In this Study we have chosen to employ and compare three different architectures to see how they fare with this dataset. The architectures chosen are ResNet50, VGG16, and MobileNetV3 Large. Each of these architectures brought unique strengths to the experiment, allowing for a comprehensive evaluation of their effectiveness in facial expression classification.

ResNet50 is a deep neural network with 50 layers designed to make training very deep models easier. It does this using "residual blocks," which have shortcut connections that skip over some layers. These shortcuts help the network learn the changes needed rather than trying to figure out the entire output from scratch. By doing so, ResNet50 manages to keep important details and gradients intact, which makes training faster and more effective. The architecture includes layers for convolution, batch normalization, and activation, all working together with these shortcuts to improve performance and training efficiency. This model was selected for its robust performance in handling complex image features [18]. The structure can be seen on Figure 3.

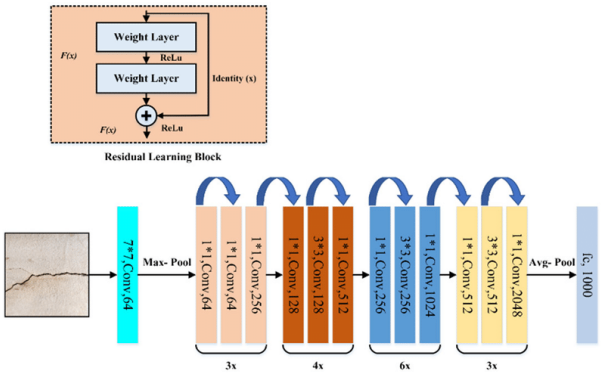

Figure 3. ResNet 50 Architecture [18]

VGG16 is a deep neural network with 16 layers, known for its simplicity and effectiveness and chosen for its simplicity and proven effectiveness in image classification,

especially when dealing with smaller datasets like the one used in this experiment. It consists mainly of stacked convolutional layers with small 3x3 filters, which helps the network learn fine details in the data. The architecture is straightforward: a series of these convolutional layers are followed by max-pooling layers to reduce the spatial dimensions. After the convolutional part, the network uses fully connected layers to make predictions. This design makes VGG16 easy to understand and implement, and it's effective at capturing high-level features from images [18]. The structure can be seen on Figure 4.

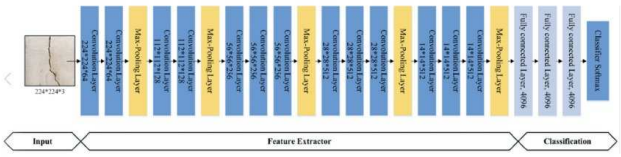

Figure 4. VGG16 Architecture [1 8]

MobileNetV3 is a compact and efficient deep neural network developed by Google, designed specifically for use on mobile devices and other embedded systems. Building on the first two versions, V1 and V2, MobileNetV3 introduces several enhancements that improve both its performance and speed. Its core innovation is the use of separable convolution operations, which help balance training accuracy with network efficiency. This design makes MobileNetV3 an excellent choice for lightweight models, delivering impressive performance while maintaining a small footprint suitable for mobile and edge computing applications. MobileNetV3 Large was employed for its efficiency and high performance on resource-constrained devices [19]. The structure can be seen on Figure 5.

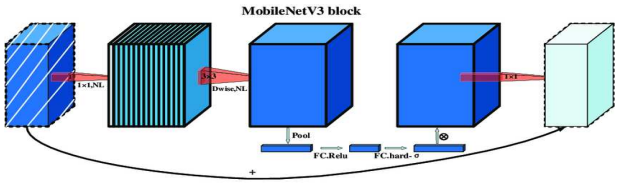

Figure 5. MobileNetV3 Large Architecture [1 9]

Transfer learning was applied to the models to leverage pre-trained weights from large-scale datasets, with three different approaches tested to assess their impact on the classification of facial expressions. The first approach involved no transfer learning, where the models were trained entirely from scratch on the dataset. The second approach employed 30% transfer learning, in which only a portion of the model—specifically, the last few layers—was fine-tuned while the earlier layers were kept frozen, allowing the model to retain learned features from the pre-trained dataset while adapting to the new task. The final approach utilized 100% transfer learning, where the entire model was fine-tuned, enabling the pre-trained features to be fully adapted to the nuances of the facial expression dataset. These varying degrees of transfer learning provided insights into how much pre-existing knowledge from large-scale datasets can be beneficial for the specific task of facial expression classification.

Utilizing 100% transfer learning allows the model to take full advantage of pre-trained weights from large-scale datasets like ImageNet. These pre-trained models have been shown to perform well in tasks like FER, where identifying general features such as facial landmarks, edges, and textures is crucial. According to [4], transfer learning accelerates model convergence and improves generalization, especially in tasks requiring similar low-level visual feature recognition.

However, 30% transfer learning involves fine-tuning part of the model, which enables it to adjust to the specific nuances of facial expression data. This technique strikes a balance between leveraging the pre-trained features and allowing the model to specialize in detecting subtle emotion-related variations that might not be present in a generalized dataset. This research [20] found that partial fine-tuning can significantly improve performance by adapting higher-level layers to more task-specific patterns, like those required for recognizing emotions in FER datasets.

Lastly, 0% transfer learning, or training from scratch, is beneficial when the target dataset is distinct from the source dataset used in pre-training, or when the dataset is rich enough to allow the model to learn domain-specific features without relying on pre-trained knowledge. As [20] pointed out, training from scratch provides flexibility, allowing the model to develop unique features for facial expression recognition without being constrained by pre-trained weights that were designed for different tasks.

By experimenting with these three levels of transfer learning, I aimed to evaluate the impact of pre-trained knowledge on FER while also allowing room for specialization and task-specific adaptation. This approach is consistent with recent research suggesting that varying levels of transfer learning can yield different results depending on the dataset's size and specificity.

The training process was conducted on a laptop equipped with an AMD Ryzen 5 5600H processor, NVIDIA® GeForce RTX™ 3060 Laptop GPU, and 16 GB of DDR4-3200 RAM, providing ample computational power for the task. Each model, across all preprocessing and transfer learning variations, was trained for 100 epochs. Hyperparameters such as learning rate and batch size were carefully optimized to achieve the best performance for each configuration. The activation function employed was ReLU, optimized using Adam, and the loss function used was categorical_crossentropy, suitable for datasets with more than one label. The training process involved continuous monitoring of loss and accuracy metrics to minimize overfitting and ensure that the models could generalize effectively to unseen data. This comprehensive approach allowed for a thorough evaluation of how different architectures, preprocessing techniques, and transfer learning strategies impact the classification of facial expressions.

III. RESULTS AND DISCUSSION

In the first experiment, no preprocessing was applied to the images. Instead, we focused on evaluating the performance of various transfer learning strategies across the selected architectures. The models were trained for 100 epochs, and the highest accuracy was observed with the VGG16 architecture without transfer learning, achieving a test accuracy of 0.70.

The second experiment involved applying an undersampling technique to ensure each class had an equal number of images. The resulting balanced dataset is depicted in Figure 3. However, this approach did not significantly improve the model's performance. The best result with the undersampled dataset was again achieved with the VGG16 architecture without transfer learning, yielding a test accuracy of 0.70. The undersampled dataset can be seen on Figure 6.

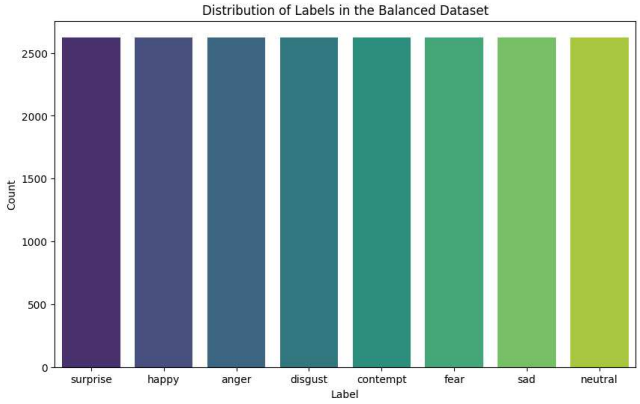

Figure 6. The undersampled dataset

For the third experiment, we employed an oversampling technique using SMOTE, combined with Contrast Limited Adaptive Histogram Equalization (CLAHE) to optimize the dataset further. The enhanced dataset following CLAHE and SMOTE preprocessing is shown in Figure 4. This preprocessing strategy resulted in a noticeable improvement in model accuracy, with the VGG16 architecture without transfer learning reaching a test accuracy of 0.79. The CLAHE+SMOTE dataset can be seen on Figure 7.

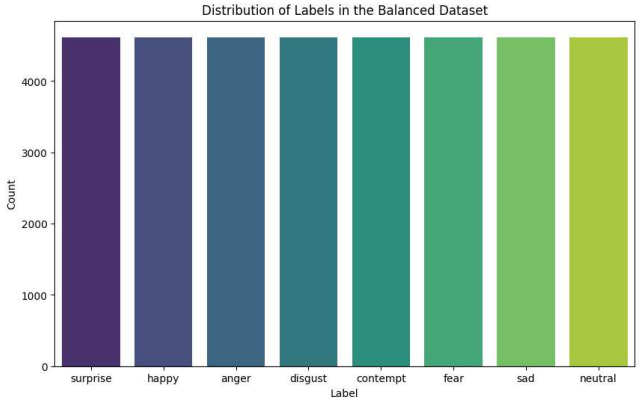

Figure 7. The CLAHE + SMOTE dataset

The comprehensive results of all experiments are summarized in Table 1.

TABLE I
PREDICTION RESULTS

Model	Training Accuracy	Validation Accuracy	Testing Accuracy	Precision	Recall	F1
ResNet50	0.95	0.68	0.67	0.65	0.65	0.64
Partially Frozen ResNet50 (30%)	0.99	0.56	0.55	0.55	0.54	0.54
Pre-Trained ResNet50	0.29	0.28	0.30	0.31	0.26	0.22
VGG16	0.99	0.72	0.70	0.72	0.72	0.72
Partially Frozen VGG16 (30%)	1.00	0.71	0.69	0.39	0.38	0.37
Pre-Trained VGG16	0.39	0.39	0.38	0.38	0.38	0.37
MobileNetV3Large	0.99	0.58	0.60	0.50	0.50	0.50
Partially Frozen MobileNetV3Large (30%)	0.98	0.50	0.48	0.41	0.40	0.40
Pre-Trained MobileNetV3Large	0.29	0.30	0.30	0.28	0.28	0.26
ResNet50 + Undersampling	0.99	0.64	0.63	0.65	0.64	0.64

Partially Frozen ResNet50 (30%) + Undersampling	0.99	0.55	0.52	0.53	0.51	0.50
Pre-Trained ResNet50 + Undersampling	0.30	0.29	0.29	0.28	0.27	0.23
VGG16 + Undersampling	1.00	0.71	0.70	0.70	0.71	0.70
Partially Frozen VGG16 (30%) + Undersampling	1.00	0.71	0.69	0.71	0.72	0.71
Pre-Trained VGG16 + Undersampling	0.42	0.39	0.39	0.39	0.38	0.38
MobileNetV3Large + Undersampling	0.99	0.50	0.52	0.49	0.49	0.49
Partially Frozen MobileNetV3Large (30%) + Undersampling	0.99	0.47	0.45	0.42	0.42	0.41
Pre-Trained MobileNetV3Large + Undersampling	0.30	0.29	0.30	0.26	0.26	0.23
ResNet50 + CLAHE + SMOTE	1.00	0.77	0.75	0.75	0.74	0.74
Partially Frozen ResNet50 (30%) + CLAHE + SMOTE	1.00	0.71	0.68	0.69	0.68	0.68
Pre-Trained ResNet50 + CLAHE + SMOTE	0.31	0.31	0.30	0.30	0.28	0.24
VGG16 + CLAHE + SMOTE	**1.00**	**0.80**	**0.79**	**0.79**	**0.79**	**0.79**
Partially Frozen VGG16 (30%) + CLAHE + SMOTE	1.00	0.79	0.79	0.78	0.78	0.78
Pre-Trained VGG16 + CLAHE + SMOTE	0.43	0.42	0.40	0.34	0.34	0.32
MobileNetV3Large + CLAHE + SMOTE	0.99	0.69	0.67	0.68	0.68	0.67
Partially Frozen MobileNetV3Large (30%) + CLAHE + SMOTE	0.99	0.65	0.62	0.62	0.62	0.62
Pre-Trained MobileNetV3Large + CLAHE + SMOTE	0.33	0.33	0.32	0.27	0.25	0.23

The application of CLAHE combined with SMOTE for dataset preprocessing proved to be effective in improving the model's accuracy across all configurations, whether using full transfer learning, partial transfer learning, or no transfer learning at all. To further evaluate and interpret the performance of the VGG16 model with CLAHE and SMOTE preprocessing, a confusion matrix is introduced. The confusion matrix on the test of dataset we used can be seen on Figure 8.

Figure 8 displays the performance of the model on the test dataset, where it achieved an overall accuracy of 0.79, with precision, recall, and F1 score also measured at 0.79. The confusion matrix highlights the model's ability to classify emotions with a good degree of accuracy across most categories. For instance, the "happiness" class was identified correctly 413 times, while the "fear" and "sadness" categories also exhibited relatively high correct predictions of 382 and 381, respectively. However, there are instances of misclassification, such as "neutral" being confused with "sadness" and "disgust" being mistaken for "neutral" and "anger." These errors indicate that while the model performs well overall, there is room for improvement in distinguishing between subtle emotional expressions.

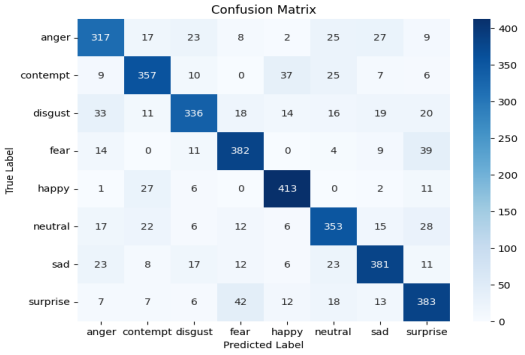

Figure 8. The confusion matrix of VGG16+CLAHE+SMOTE

In addition to testing on the initial dataset, the model was also evaluated on the FER2013Plus dataset to further validate its generalizability across different datasets. The FER2013Plus dataset contains a variety of facial expressions, adding robustness to the evaluation of the VGG16 model with CLAHE and SMOTE preprocessing. The confusion matrix for the FER2013Plus dataset is presented in Figure 9.

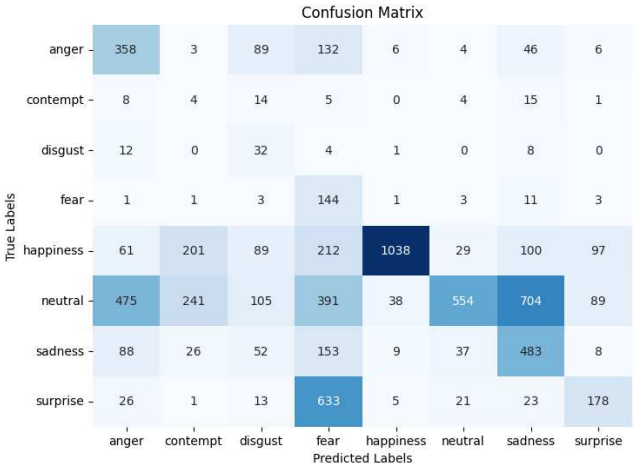

Figure 9. The confusion matrix model on FER2013Plus

In addition to testing on the initial dataset, the model was also evaluated on the FER2013Plus dataset to further validate its generalizability across different datasets. The FER2013Plus dataset contains a variety of facial expressions, adding robustness to the evaluation of the VGG16 model with CLAHE and SMOTE preprocessing. The confusion matrix for the FER2013Plus dataset is presented in Figure 9.

The model exhibits strong performance in classifying the happiness class, with 1,038 correct classifications. This high number suggests that happiness, being a more distinct and easily recognizable emotion, is consistently identified by the model. However, the performance drops noticeably in other categories, revealing specific areas of weakness in the model's ability to differentiate between similar emotional expressions.

For instance, the neutral class is one of the most misclassified categories. The model incorrectly labels neutral expressions as happiness (391 instances) and sadness (704 instances). This misclassification could arise from the inherent subtlety of neutral expressions, which often share facial features with other emotions. The overlap in facial cues between neutral, happiness, and sadness seems to confuse the model, causing significant errors in classification.

Another notable area of misclassification is in the surprise class, where the model confuses surprise with happiness in 633 instances. This confusion might stem from the fact that surprise and happiness can share some visual similarities in facial expressions, particularly around the mouth and eyes, depending on the intensity of the surprise or excitement.

The sadness class, though not as frequently misclassified as neutral, still shows significant errors, with 483 correct predictions and a substantial number of misclassifications into neutral (153 instances). This indicates that the model finds it challenging to differentiate between these two emotions, possibly because sadness and neutral expressions can overlap in subtle facial cues when the intensity of the sadness is mild.

Other emotions, such as anger and fear, show a moderate number of correct classifications but still suffer from a degree of confusion with neutral and happiness. For example, anger is confused with neutral (475 instances), which further points to the model's difficulty in distinguishing between strong emotions and more subdued or passive states like neutrality.

Overall, the confusion matrix from Figure 9 indicates that while the model has a strong ability to identify more distinct emotions like happiness, it struggles with more subtle or overlapping emotions, such as neutral, sadness, and surprise. The overall precision 0.68, recall 0.39, and F1 score 0.44 reflect this imbalance, showing that while the model can make precise predictions for some classes, its recall is limited, meaning that many true instances of certain emotions (like neutral) are being missed or misclassified.

IV. CONCLUSION

This study investigated the impact of various preprocessing techniques and deep learning architectures on the classification of facial expressions using the "Affects" dataset. The dataset, which presented challenges such as class imbalance and varying image quality, was processed using methods like undersampling, CLAHE, and SMOTE to enhance model performance.

The experiments demonstrated that the combination of CLAHE and SMOTE preprocessing was particularly effective in improving the accuracy of the models. While initial experiments using the original and undersampled dataset produced moderate results, with a maximum test accuracy of 0.70 respectively, the application of CLAHE and SMOTE led to a significant boost in classification performance. The VGG16 architecture, without the use of transfer learning, achieved a test accuracy of 0.79 when trained on the CLAHE + SMOTE preprocessed dataset, representing an improvement of approximately 0.09 or 9% over the best result obtained without these preprocessing techniques.

These findings highlight the critical role of advanced preprocessing techniques in handling imbalanced and low-contrast datasets. The use of CLAHE enhanced the visibility of facial features by improving image contrast, while SMOTE effectively balanced the dataset by generating synthetic samples for minority classes. This combination not only improved model accuracy by nearly 9% but also demonstrated the potential of preprocessing to make a more substantial impact than architectural changes alone.

In conclusion, the success of CLAHE and SMOTE in enhancing the performance of facial expression recognition models underscores the importance of tailored preprocessing in deep learning workflows. Future research could further

explore the refinement of these techniques and their application to other domains, paving the way for even more robust and accurate models.

AUTHORS CONTRIBUTION

Bryan Anthony: Investigation, Original Draft Writing Preparation, Project Administration, Resources, Software, Visualization, Original Drafting Writing;
Nicholas Dylan Lienardi: Investigation, Software Visualization, Writing Original Draft Writing Preparation;
Richard Evan Sutanto: Formal Analysis, Conceptualization, Investigation, Methodology, Supervision, Validation, Review Writing & Editing;
Yuwono Marta Dinata: Project Administration, Review & Editing, Writing Original Draft Writing Preparation;

COPYRIGHT

REFERENCES

[1] T.-H. Vo, G.-S. Lee, H.-J. Yang, and S.-H. Kim, "Pyramid with super resolution for in-the-wild facial expression recognition," *IEEE Access*, vol. 8, pp. 131988–132001, 2020, doi: 10.1109/access.2020.3010018.
[2] G. Gaudi, B. Kapralos, K. C. Collins, and A. Quevedo, "Affective computing: An introduction to the detection, measurement, and current applications," in *Learning and Analytics in Intelligent Systems*, pp. 25–43, 2021, doi: 10.1007/978-3-030-80571-5_3.
[3] M. Sajjad, S. Zahir, A. Ullah, Z. Akhtar, and K. Muhammad, "Human behavior understanding in big multimedia data using CNN-based facial expression recognition," *Mobile Networks and Applications*, vol. 25, no. 4, pp. 1611–1621, 2020, doi: 10.1007/s11036-019-01366-9.
[4] S. Minaee, M. Minaei, and A. Abdolrashidi, "Deep-emotion: Facial expression recognition using attentional convolutional network," *Sensors*, vol. 21, no. 9, p. 3046, 2021, doi: 10.3390/s21093046.
[5] T. Kalsum, S. M. Anwar, M. Majid, B. Khan, and S. M. Ali, "Emotion recognition from facial expressions using hybrid feature descriptors," *IET Image Processing*, vol. 12, no. 6, pp. 1004–1012, 2018, doi: 10.1049/iet-ipr.2017.0499.
[6] X. Zhu and Y. Wang, "Hybrid attention cascade network for facial expression recognition," *Journal of Big Data*, vol. 7, no. 1, p. 46, 2020, doi: 10.1186/s40537-020-00317-8.
[7] G. Oh, Y. Kim, S. Kim, S. Choi, and J. Park, "DRER: Deep learning–based driver's real emotion recognizer," *Sensors*, vol. 21, no. 6, p. 2166, 2021, doi: 10.3390/s21062166.
[8] A. Chaudhari, C. Bhatt, A. Krishna, and P. L. Mazzeo, "ViTFER: Facial emotion recognition with vision transformers," *Applied System Innovation*, vol. 5, no. 4, p. 80, 2022, doi: 10.3390/asi5040080.
[9] S. S. Hiremath, J. Hiremath, V. V. Kulkarni, B. C. Harshit, S. Kumar, and M. S. Hiremath, "Facial expression recognition using transfer learning with ResNet50," *Lecture Notes in Networks and Systems*, pp. 281–300, 2023, doi: 10.1007/978-981-99-1624-5_21.
[10] O. Khajuria, R. Kumar, and M. Gupta, "Facial emotion recognition using CNN and VGG-16," in *2023 International Conference on Inventive Computation Technologies (ICICT)*, pp. 472–477, 2023, doi: 10.1109/icict57646.2023.10133972.
[11] R. S. Alshwihde and W. I. Eltarhouni, "Enhanced facial expression recognition using pre-trained models and image processing techniques," *Communications in Computer and Information Science*, pp. 269–283, 2024, doi: 10.1007/978-3-031-62624-1_22.
[12] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "AffectNet: A database for facial expression, valence, and arousal computing in the wild," *IEEE Transactions on Affective Computing*, vol. 10, no. 1, pp. 18–31, 2017, doi: 10.1109/TAFFC.2017.2740923.
[13] K. Zuiderveld, "Contrast limited adaptive histogram equalization," in *Graphics Gems*, pp. 474–485, 1994, doi: 10.1016/b978-0-12-336156-1.50061-6.
[14] F. I. Ilmawati, K. Kusrini, and T. Hidayat, "Optimizing facial expression recognition with image augmentation techniques: VGG19 approach on FERC dataset," *Sinkron*, vol. 8, no. 2, pp. 632–640, 2024, doi: 10.33395/sinkron.v8i2.13507.
[15] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002, doi: 10.1613/jair.953.
[16] Y. Tian, T. Han, and L. Wu, "Teacher facial expression recognition based on GoogLeNet-InceptionV3 CNN model," in *Artificial Intelligence in Education and Teaching Assessment*, pp. 69–78, 2021, doi: 10.1007/978-981-16-6502-8_8.
[17] H. He and E. Garcia, "Learning from imbalanced data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 9, pp. 1263–1284, 2009, doi: 10.1109/tkde.2008.239.
[18] L. Ali, F. Alnajjar, H. A. Jassmi, M. Gocho, W. Khan, and M. A. Serhani, "Performance evaluation of deep CNN-based crack detection and localization techniques for concrete structures," *Sensors*, vol. 21, no. 5, p. 1688, 2021, doi: 10.3390/s21051688.
[19] Y. Wu, Y. Sun, S. Zhang, X. Liu, K. Zhou, and J. Hou, "A size-grading method of antler mushrooms using YOLOv5 and PSPNet," *Agronomy*, vol. 12, no. 11, p. 2601, 2022, doi: 10.3390/agronomy12112601.
[20] M. Mohana and P. Subashini, "Facial expression recognition using machine learning and deep learning techniques: A systematic review," *SN Computer Science*, vol. 5, no. 4, 2024, doi: 10.1007/s42979-024-02792-7.