Deep Learning-Based Facial Expression Recognition for Analyzing Visitor Engagement

Grace Setiaputri, Jason Chrisbellno Mackenzie, Ivan Sebastian Edbert and Derwin Suhartono
Computer Science, School of Computer Science, Bina Nusantara University, Jakarta, 11480, Indonesia

Keywords: Face Expression, ResNet50, VGG, EfficientNetB3.

Abstract: Interpersonal communication relates to facial expression because it's a natural source. In the real world, people use their facial expressions, particularly when conducting customer satisfaction surveys for business purposes. Due to the increasing number of fake reviews, evaluating customer satisfaction based on online reviews is sometimes inaccurate. This study uses three machine learning models, ResNet50, VGG16, and EfficientNetB3, to classify human facial expressions. The FER-2013 dataset is used, then oversampled and augmented, the performance of three models was compared using accuracy and F1-Score as comparison values. EfficientNetB3 gets the highest accuracy of 85.41% and F1-score of 85.34%. Future research should apply more sophisticated data balancing techniques, such as the Synthetic Minority Over-sampling Technique (SMOTE), to address data imbalances without adding processing time. Furthermore, extending the number of epochs and refraining from early stopping strategies may help in determining the model's maximal accuracy potential.

1 INTRODUCTION

Interpersonal communication depends heavily on facial expressions, which are a way of visualizing the emotions that people experience (Xiong et al., 2024). People can communicate their feelings honestly and directly through facial expressions. For instance, a grin can convey joy, but a scowl might convey anxiety or discontent.

Business owners typically gauge client satisfaction by reading online reviews. However, this method usually require to provide precise and reliable insights. The percentage of phony reviews that consumers will come across in 2022, 2023, and 2024 across various media is displayed in Figure 1. According to the 2024 Local Consumer Review Survey by Bright Research, 82% of false reviews are spread across multiple platforms, which could lead to bias evaluating consumer satisfaction (Why Are Fake Online Reviews a Problem? - BrightLocal, n.d.-a). A business's ability to evaluate customer happiness is significantly impacted. For instance, a 5-9% boost in sales can be achieved by improving a restaurant's Yelp rating (Customer Reviews: Stats That Demonstrate the Impact of Reviews - ReviewTrackers, n.d.). Consequently, business owners need to utilize face expression analysis.

AI and machine learning (ML) have driven numerous studies in facial expression estimation. Approaches from 2021-2024 include CNN with ResMaskingNet and bResNet-18 (L. Pham et al., 2021a), (Chaudhari et al., 2022a) . Other 2024 research utilized CNN ensemble models (Xception, DCNN, VGG16 (Melinte & Vladareanu, 2020a), EfficientNetB2, DenseNet (Dong et al., 2023a), RestNet50, and InceptionRestnetv2) and additional ensemble models (AlexNet, Inception V3, ResNet50) (Lawpanom et al., 2024a), (Reghunathan et al., 2024a). Despite using datasets like FER-2013, these models often struggle to fully capture facial expressions due to insufficient depth for complex features, poor generalization, and vanishing gradient issues.

In its application later when using deep learning, data in the form of images of expressions obtained when visitors visit the place will be used as a reference for assessing visitor satisfaction as well as being used as material for business evaluation.
This research aims to identify the most effective facial expression recognition model among ResNet50, EfficientNetB3, and VGG16, using the FER-2013 dataset. The goal is to help business

Setiaputri, G., Chrisbellno Mackenzie, J., Sebastian Edbert, I. and Suhartono, D.
Deep Learning-Based Facial Expression Recognition for Analyzing Visitor Engagement.
DOI: 10.5220/0014268200004928

In *Proceedings of the 1st International Conference on Research and Innovations in Information and Engineering Technology (RITECH 2025)*, pages 32-36
ISBN: 978-989-758-784-9

owners improve service quality through better understanding of customer expressions.

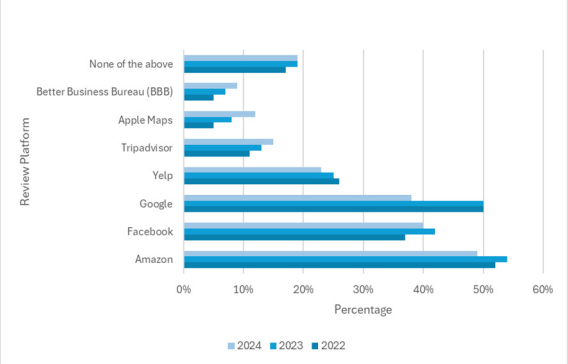

Figure 1: Percentage of fake reviews on each review platform (*Why Are Fake Online Reviews a Problem? - BrightLocal*, n.d.-b).

2 LITERATURE REVIEW

Facial expressions are crucial for conveying human emotions (Xiong et al., 2024), making their accurate interpretation valuable for businesses. While various models like Random Forest, CNN, Residual Masking Network (L. Pham et al., 2021b; Yeh et al., 2023), CNN with C4.5, Vision Transformers (ViT) (Dosovitskiy et al., 2020), and ensemble models have been explored (Chaudhari et al., 2022b; Lawpanom et al., 2024b; Wang et al., 2019a). Their accuracy in facial expression recognition has been less than satisfactory. This research aims to identify the most effective method for this task.

Several studies use the FER-2013 dataset provided by Kaggle (Lawpanom et al., 2024b; Yeh et al., 2023). Other studies provide additional datasets, such as JAFFE, CK+, and RAF-DB, to test the effectiveness of various algorithms and techniques (Wang et al., 2019b). The VEMO, AffectNet, and CK+48 datasets are also leveraged to improve model robustness and generalize performance over a wide range of facial expressions (Chaudhari et al., 2022b; L. Pham et al., 2021b).

In 2019, researchers used the Random Forest Algorithm to detect facial expressions. The study uses the C4.5 classifier, widely applied in image recognition due to its vast data handling capacity, ideal computing efficiency, and intuitive features (Wang et al., 2019a). This study uses FER-2013 datasets and produces 84.3% accuracy with a running time of 15,860.5(s).

A 2021 study introduced a Residual Masking Network that uses a Unet-based localization network for masking. This approach allows the model to concentrate on crucial spatial information by employing a segmentation network to refine feature maps. Even with imbalanced data, the model achieved strong performance on the FER-2013 dataset (L. Pham et al., 2021b).

In 2024, a study highlights that ensemble deep learning models surpass individual ones. This research utilizes the HoE-7 ensemble, comprising Xception, DCNN, VGG16 (Melinte & Vladareanu, 2020b), EffcientNetB2, DenseNet (Dong et al., 2023b), RestNet50, and InceptionResNetV2, achieving 75.15% accuracy. The HoE-6 ensemble obtained a slightly lower 73.73% accuracy (Lawpanom et al., 2024b). This superior performance is attributed to ensembles' ability to mitigate overfitting, reduce prediction variance and bias, and compensate for individual model errors.

3 METHODOLOGY

This study will recognize face expressions using deep learning model, namely CNN with Resnet50 and EfficientNetB3 architectures (Zhu et al., 2022), as well as VGG (Cheng & Kong, 2024). The methodology starts with data collection, preprocessing, and splitting so that it is ready for each model. It is followed by model building and hyperparametric tuning. The results will be tested and compared to obtain the best-performing model for classifying facial expression recognition.

3.1 Data Collection

This research uses a publicly available dataset from the Kaggle platform. Kaggle created this dataset for facial expression recognition research (FER-2013, n.d.) as shown in Figure 2. There are 32,298 photos grayscale images depicting seven facial expressions: happy, anger, fear, surprise, sad, disgust, and neutral. Based on this dataset, this research will classify facial expression using various models.

3.2 Data Preprocessing

Preprocessing data is a essential process, aimed at converting unprocessed data into a structured format optimized for machine learning applications (Zhao et al., 2021). Data preprocessing techniques will be

carried out in this stage: rescaling, resizing, and label encoding.

Figure 2: Example Expression from FER-2013 Dataset.

3.2.1 Rescaling

Before processing, each loaded image will be rescaled based on its pixel value by dividing by 255 so that the data can be in the range 0 to 1 (Zhao et al., 2021).

3.2.2 Resizing

Each processed image is resized to 48x48 pixels (*FER-2013*, n.d.). This uniformity ensures that the model can efficiently learn from the images without being affected by varying image sizes.

3.2.3 Label Encoding

One-hot encoding converts categorical features into numerical ones by creating a new binary feature for each category, indicating its presence (1) or absence (0).

3.3 Data Splitting

Splitting the dataset into training and testing sets is crucial to enhance the generalizability and robustness of the facial expression recognition model (Dong et al., 2023b). This research used in facial expression recognition includes 80% for training and 20% for testing.

3.4 Model Training

Three deep learning models have been compared to classify human facial expressions to determine visitors' interest in a place. Those three models are described below.

3.4.1 Resnet50

A deep residual learning architecture is well known for facial expression recognition tasks due to its effectiveness in learning image features (Tsalera et al., 2022). This model comprises 50 layers, featuring 16 residual blocks with three layers each and input and output layers (Reghunathan et al., 2024b). The structure combines deep learning with residual learning, which can create connections so that information flows from one layer to another (Chouhayebi et al., 2024).

3.4.2 EfficientNetB3

This architecture is widely applied to facial expression detection, improving overall recognition performance by identifying and extracting attributes from facial images. EfficientNet has seven multidimensional models—which use scaling and AutoML—the model performs better than the majority of convolutional neural networks (Zhu et al., 2022).

3.4.3 VGG16

A popular deep convolutional neural network design for image classification applications, such as face expression detection, is VGG16 (Chouhayebi et al., 2024). VGG16 contains 138,355,752 parameters, three deep layers, and five convolution blocks. Figure 3 shows the architecture of VGG used. Block output size and noise are decreased via convolutional layers in conjunction with a max pooling layer (Zhu et al., 2022).

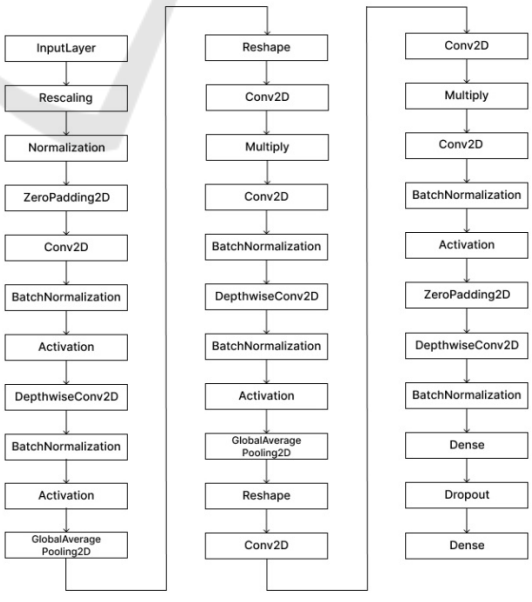

Figure 3: Architectures of VGG16 (Melinte & Vladareanu, 2020b).

3.5 Hyperparametric Tuning

To optimize facial expression interpretation, model parameters will be tuned using the Hyperband method from Keras Tuner. This study will focus on refining the learning rate, batch size, dense unit, optimizer, and regularization for each model.

3.6 Evaluation

Recall, accuracy (Melinte & Vladareanu, 2020b), precision, and F1 score are key evaluation criteria (Yeh et al., 2023), each offering insights into a model's performance by highlighting its strengths and weaknesses.

4 RESULT AND DISCUSSION

This study evaluated the classification performance of ResNet50, EfficientNetB3, and VGG16 models on the FER-2013 dataset. Model accuracy was assessed using confusion matrices and classification reports, which included F1 score, recall, and precision. As shown in Table 1, the EfficientNetB3 model achieved the highest accuracy score of 0.84467, outperforming the other models. Its consistent precision, recall, and F1-score values further indicate that EfficientNetB3 provides more reliable predictions compared to ResNet50 and VGG16.

Table 1: Model Comparison.

Model	Accuracy	Precision	Recall	F1
Efficient NetB3	0.85417	0.85421	0.85417	0.85347
ResNet50	0.76794	0.7673	0.76794	0.76731
VGG16	0.76626	0.76546	0.76626	0.76545

The EfficientNetB3 model demonstrates strong performance in classifying facial expressions, as detailed in its confusion matrix (Figure 4). This is because EfficientNetB3 uses compound scaling, which increases the model's capacity not only from depth, width, or input resolution alone, but by increasing all of them in a balanced way. It achieved exceptional accuracy for "Disgust" and "Surprise," with 1443 and 1395 correct predictions respectively, and minimal misclassifications across other categories. The "Angry" expression also showed good accuracy with 1261 true positives. "Fear" and "Neutral" expressions were classified accurately 1220 and 1099 times, respectively. While "Happy" and "Sad" emotions proved more challenging to differentiate, the model still yielded 1171 and 1039 true positives for these classes. Overall, the consistent results across all classes indicate that the EfficientNetB3 model effectively captures and recognizes crucial features of each facial expression with stable accuracy.

EfficientNetB3 demonstrates superior facial expression classification compared to ResNet50 and VGG16. This conclusion is drawn from the classification report and confusion matrix, which further indicate a balanced distribution of true positives across all classes, signifying no significant data imbalance.

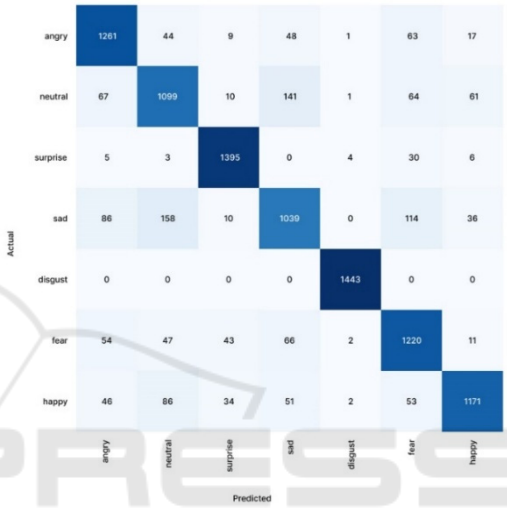

Figure 4: Confusion Matrix for EfficientNetB3.

5 CONCLUSIONS

Deep learning facial recognition uses human facial characteristics to help businesses manage and improve services. A study compared ResNet50, EfficientNetB3, and VGG16 models for facial expression recognition using the FER-2013 dataset.[2] After data preprocessing, EfficientNetB3 achieved the highest accuracy at 0.84467, outperforming ResNet50 and VGG16. All models provide reliable predictions, with this method showing particular strength in classifying "Disgust" and "Surprise" expressions.

This study's limitations stem from its extensive and irregular data, particularly imbalances between expression classes. While oversampling addressed this, it increased preprocessing complexity and total processing time. Consequently, model performance remained slightly suboptimal in classes with sparser data. In addition, the use of a single dataset such as FER-2013 poses a challenge for the model in the generalization process.

For future research, it's recommended to employ more sophisticated balancing techniques, such as Synthetic Minority Oversampling Technique (SMOTE), the use of SMOTE has a very positive impact on balancing classes by creating new synthetic data for minority classes. Additionally, to maximize the level of accuracy, adding epochs and loss function (T.-D. Pham et al., 2023) is the right step so that the model can capture the complexity of the data and low validation loss. Furthermore, using other datasets such as RAF-DB, FERPlus (Yao et al., 2023), AffectNet, or CK+ is highly recommended for cross-validation and proving that the model is robust or generalizable.

REFERENCES

Chaudhari, A., Bhatt, C., Krishna, A., & Mazzeo, P. L. (2022a). ViTFER: Facial Emotion Recognition with Vision Transformers. *Applied System Innovation*, *5*(4), 80. https://doi.org/10.3390/asi5040080

Cheng, Y., & Kong, D. (2024). CSINet: Channel–Spatial Fusion Networks for Asymmetric Facial Expression Recognition. *Symmetry*, *16*(4), 471. https://doi.org/10.3390/sym16040471

Chouhayebi, H., Mahraz, M. A., Riffi, J., Tairi, H., & Alioua, N. (2024). Human Emotion Recognition Based on Spatio-Temporal Facial Features Using HOG-HOF and VGG-LSTM. *Computers*, *13*(4), 101. https://doi.org/10.3390/computers13040101

Customer Reviews: Stats that Demonstrate the Impact of Reviews - ReviewTrackers. (n.d.). Retrieved May 29, 2024, from https://www.reviewtrackers.com/reports/customer-reviews-stats/

Dong, J., Zhang, Y., & Fan, L. (2023a). A Multi-View Face Expression Recognition Method Based on DenseNet and GAN. *Electronics*, *12*(11), 2527. https://doi.org/10.3390/electronics12112527

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2020). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *ICLR 2021 - 9th International Conference on Learning Representations*. https://arxiv.org/abs/2010.11929v2

FER-2013. (n.d.). Retrieved June 10, 2024, from https://www.kaggle.com/datasets/msambare/fer2013/data

Lawpanom, R., Songpan, W., & Kaewyotha, J. (2024a). Advancing Facial Expression Recognition in Online Learning Education Using a Homogeneous Ensemble Convolutional Neural Network Approach. *Applied Sciences*, *14*(3), 1156. https://doi.org/10.3390/app14031156

Melinte, D. O., & Vladareanu, L. (2020a). Facial Expressions Recognition for Human–Robot Interaction Using Deep Convolutional Neural Networks with Rectified Adam Optimizer. *Sensors*, *20*(8), 2393. https://doi.org/10.3390/s20082393

Pham, L., Vu, T. H., & Tran, T. A. (2021b). Facial Expression Recognition Using Residual Masking Network. *2020 25th International Conference on Pattern Recognition (ICPR)*, 4513–4519. https://doi.org/10.1109/ICPR48806.2021.9411919

Pham, T.-D., Duong, M.-T., Ho, Q.-T., Lee, S., & Hong, M.-C. (2023). CNN-Based Facial Expression Recognition with Simultaneous Consideration of Inter-Class and Intra-Class Variations. *Sensors*, *23*(24), 9658. https://doi.org/10.3390/s23249658

Reghunathan, R. K., Ramankutty, V. K., Kallingal, A., & Vinod, V. (2024b). Facial Expression Recognition Using Pre-trained Architectures. *The 2nd Computing Congress 2023*, 22. https://doi.org/10.3390/engproc2024062022

Tsalera, E., Papadakis, A., Samarakou, M., & Voyiatzis, I. (2022). Feature Extraction with Handcrafted Methods and Convolutional Neural Networks for Facial Emotion Recognition. *Applied Sciences*, *12*(17), 8455. https://doi.org/10.3390/app12178455

Wang, Y., Li, Y., Song, Y., & Rong, X. (2019b). Facial Expression Recognition Based on Random Forest and Convolutional Neural Network. *Information*, *10*(12), 375. https://doi.org/10.3390/info10120375

Why Are Fake Online Reviews a Problem? - BrightLocal. (n.d.-a). Retrieved May 29, 2024, from https://www.brightlocal.com/learn/review-management/fake-reviews/why-are-fake-reviews-a-problem/

Transformer and Adaptive Method with Visual Transformer for Robust Facial Expression Recognition. *Applied Sciences*, *14*(4), 1535. https://doi.org/10.3390/app14041535

Yao, H., Yang, X., Chen, D., Wang, Z., & Tian, Y. (2023). Facial Expression Recognition Based on Fine-Tuned Channel–Spatial Attention Transformer. *Sensors*, *23*(15), 6799. https://doi.org/10.3390/s23156799

Yeh, J.-F., Lin, K.-M., Chang, C.-C., & Wang, T.-H. (2023). Expression Recognition of Multiple Faces Using a Convolution Neural Network Combining the Haar Cascade Classifier. *Applied Sciences*, *13*(23), 12737. https://doi.org/10.3390/app132312737

Zhao, Z., Liu, Q., & Zhou, F. (2021). Robust Lightweight Facial Expression Recognition Network with Label Distribution Training. *Proceedings of the AAAI Conference on Artificial Intelligence*, *35*(4), 3510–3519. https://doi.org/10.1609/aaai.v35i4.16465

Zhu, H., Xu, H., Ma, X., & Bian, M. (2022). Facial Expression Recognition Using Dual Path Feature Fusion and Stacked Attention. *Future Internet*, *14*(9), 258. https://doi.org/10.3390/fi14090258