KLIK: Kajian Ilmiah Informatika dan Komputer
ISSN 2723-3898 (Media Online)
Vol 4, No 6, June 2024, Hal 2676-2685
DOI 10.30865/klik.v4i6.1874
https://djournals.com/klik

Enhanced Facial Expression Recognition Through a Hybrid Deep Learning Approach Combining ResNet50 and ResNet34 Models

Sigit Auliana, Siti Mahrojah*, Gagah Dwiki Putra Aryono

Faculty of Computer Science, Information System, Universitas Bina Bangsa, Serang, Indonesia
Email: [1]pasigit@gmail.com, [2,*]sitimahrojah3@gmail.com, [3]gagahdpa@gmail.com[3]
Email Corresponding Author: sitimahrojah3@gmail.com

Abstract-Recognizing facial expressions is a critical aspect of computer vision and human-computer interaction. It facilitates the interpretation of human emotions from facial images, aiding in applications such as affective computing, social robotics, and psychological research. In this work, we propose using hybrid deep learning models, ResNet50 and ResNet34, for facial expression classification. These models, pre-trained on large-scale datasets, demonstrate exceptional feature extraction capabilities and have achieved excellent performance in various computer vision tasks. Our approach begins with the collection and preprocessing of a labeled facial expression dataset. The collected data undergoes face detection, alignment, and normalization to ensure consistency and reduce noise. After preprocessing, the dataset is divided into training, validation, and testing sets. We fine-tune the ResNet50 and ResNet34 models on the training set, employing transfer learning to adapt the pre-trained models specifically for the facial expression recognition task. Optimization techniques such as SGDM, ADAM, and RMSprop are used to update the models' parameters and minimize the categorical cross-entropy loss function. The trained models are evaluated on the validation set, achieving an accuracy of 98.19%. Subsequently, the models are tested on unseen facial images to assess their generalization capabilities. This proposed approach aims to deliver accurate and robust facial expression classification, thereby advancing emotion analysis and human-computer interaction systems.

Keywords: CNN; Attention Mechanism; ResNet50; ResNet34; FER

1. INTRODUCTION

Emotions are fundamental to human experience and play a significant role in interpersonal communication [1]. People express their emotions in various ways, including through language, body language, and facial expressions . Among these, the analysis of facial movements is the most extensively researched method for determining emotions [2]. Extensive studies by researchers have identified universal facial expressions corresponding to emotions such as happiness, sadness, anger, fear, surprise, disgust, and neutrality [3]–[5]. Recently, interpreting emotions from facial expressions has gained considerable interest in psychology, psychiatry, and mental health research [6], [7].

The automated recognition of emotions from facial expressions is essential for advancing "smart living" technologies and enhancing healthcare systems, facilitating more intuitive and responsive interactions between humans and intelligent environments. This capability is important for diagnosing emotional disorders in conditions such as autism spectrum disorder and schizophrenia, as well as for applications in human-computer interaction (HCI) and human-robot interaction (HRI), including social welfare schemes based on HRI [8], [9]. Consequently, facial emotion recognition (FER) has garnered significant attention from researchers due to its promising and diverse applications. The primary goal of FER is to map various facial expressions to their corresponding emotional states [10].

A standard Facial Emotion Recognition (FER) system comprises two primary steps: feature extraction and emotion classification. Additionally, preprocessing of images is necessary, involving tasks such as face detection, cropping, resizing, and normalization [11]. Face detection isolates the faces by removing the background and any non-facial elements. In a traditional FER system, the crucial task is extracting features from the preprocessed image [12].

Current systems employ various methods for feature extraction, including discrete wavelet transform (DWT), linear discriminant analysis, and other similar techniques [13], [14]. The extracted features are then used for emotion classification, typically using neural networks (NN) and other machine learning methods [15]–[17]. Recently, Deep Neural Networks (DNNs), particularly Convolutional Neural Networks (CNNs), have gained significant attention in FER due to their inherent ability to extract features from images [18]–[20].

Several studies have explored using CNNs to address FER problems [21], [22]. However, current FER methods often utilize CNNs with only a few layers, despite evidence that deeper models perform better in other image processing tasks [23]. This may be due to the unique challenges associated with FER. Firstly, recognizing emotions requires high-resolution images, which involve processing large amounts of data [24]. Secondly, the subtle differences between facial expressions for different emotions make classification more difficult [25].

Conversely, an extremely complex CNN comprises many concealed convolutional layers, posing difficulties during training and frequently leading to inadequate adjustment. Due to the vanishing gradient problem, increasing the number of layers beyond a certain point does not enhance accuracy [26]. To improve the accuracy of deep CNNs, various modifications and training techniques can be employed. Pre-trained deep convolutional neural network models such as VGG-16, ResNet-50, ResNet-152, Inception-v3, and DenseNet-161 are frequently used [27]–[29]. However, developing such deep models requires extensive data and significant computational power.

A seminal study in the field of emotion recognition identified six primary emotions: happiness, sadness, anger, surprise, fear, and disgust (excluding neutral) [30]. Subsequently, the development of the Facial Action Coding System (FACS) by Ekman built upon this work, establishing it as the standard for emotion recognition studies [31]. Over time,

most emotion recognition datasets also incorporated the neutral expression, expanding the total number of basic emotions to seven.

This approach was considered the most reliable at the time. Well-known hand-crafted features such as the histogram of oriented gradients (HOG) and local binary patterns (LBP) were used to identify facial emotions. Subsequently, a classifier would determine the most appropriate emotion for the image [32]. These methods performed well on simpler datasets. However, as datasets became more complex and intra-class variation increased, the limitations of these methods became apparent. For a more comprehensive grasp of potential challenges, readers may consult the visuals in the initial row of Figure 1, depicting issues such as faces being obstructed by hands or glasses, or visible only partially.Deep learning, particularly convolutional neural networks (CNNs), has achieved significant success in addressing various issues related to image classification and other vision tasks. This success has led many companies to develop FER models based on deep learning [33]–[35]. A study demonstrated the effectiveness of CNNs in accurately identifying emotions by employing a CNN without bias on the extended Cohn–Kanade dataset (CK+) and the Toronto Face Dataset (TFD), achieving state-of-the-art results in FER [36]. Additionally, deep learning techniques were employed in another study to model the facial expressions of stylized animated characters. This involved training separate networks for human and animated facial expressions, along with mapping human images to animated ones [37]. A separate research endeavor presented a neural network architecture tailored for facial expression recognition, featuring two convolutional layers, one max-pooling layer, and four inception levels (sub-networks), culminating in the creation of a five-layered network. [38]. Furthermore, a research endeavor integrated feature extraction and classification into a unified looped network, enhancing the synergy between these procedures and achieving peak accuracy on both the CK+ and JAFFE datasets [39].

While previous studies have demonstrated the efficacy of CNNs in FER, such as employing CNNs on datasets like CK+ and TFD, and even adapting networks for animated character expressions, there has been limited exploration into the synergies between different CNN architectures. To address this gap, a novel research endeavor could focus on combining the strengths of two prominent CNN architectures, ResNet50 and ResNet34, in a hybrid deep learning approach for facial expression recognition. ResNet architectures are renowned for their depth, enabling them to capture intricate features effectively. ResNet50, with its deeper architecture, can capture more complex patterns, while ResNet34, being slightly shallower, may offer computational efficiency without compromising performance significantly.

The proposed research could investigate how a hybrid model leveraging both ResNet50 and ResNet34 can enhance FER accuracy compared to using either architecture individually. This approach could involve various strategies, such as feature fusion at different layers, ensemble learning techniques, or hierarchical feature extraction. Furthermore, the research could explore the transferability of features learned from different layers of ResNet50 and ResNet34 across datasets, considering variations in facial expressions and image characteristics. Fine-tuning and transfer learning methodologies could be employed to adapt the pre-trained models to the specific task of FER.

Additionally, the study could evaluate the robustness of the hybrid approach to variations in input data quality, such as noise, occlusions, or changes in illumination conditions. Robustness analysis could involve augmenting the training data with synthetic variations or conducting experiments on diverse real-world datasets. By investigating the synergy between ResNet50 and ResNet34 architectures in a hybrid deep learning framework for FER, this research could contribute to advancing the state-of-the-art in facial expression recognition and pave the way for more effective and efficient models in real-world applications.

2. RESEARCH METHODOLOGY

This study introduces an end-to-end deep learning framework that employs an attentional convolutional network to classify emotions in facial images. The authors introduce hybrid deep learning models that combine ResNet50 and ResNet34 for facial expression classification. These pretrained models on large-scale datasets possess robust feature extraction capabilities and have shown excellent performance in various computer vision tasks. The methodology begins with the collection and preprocessing of a labeled facial expression dataset, which undergoes face detection, alignment, and normalization to ensure consistency and eliminate noise or artifacts that could impede accurate classification. The preprocessed dataset is then split into training, validation, and testing sets. The pretrained ResNet50 and ResNet34 models are fine-tuned on the training set using transfer learning, allowing the models to adapt their learned features specifically for facial expression recognition. This approach reduces training time and enhances the models' performance.

To optimize the models' performance during training, techniques such as Stochastic Gradient Descent with Momentum (SGDM), Adaptive Moment Estimation (ADAM), and Root Mean Square Propagation (RMSprop) are used [40], [41]. These methods assist in finding the optimal weights and biases, thereby enhancing the models' accuracy and convergence.

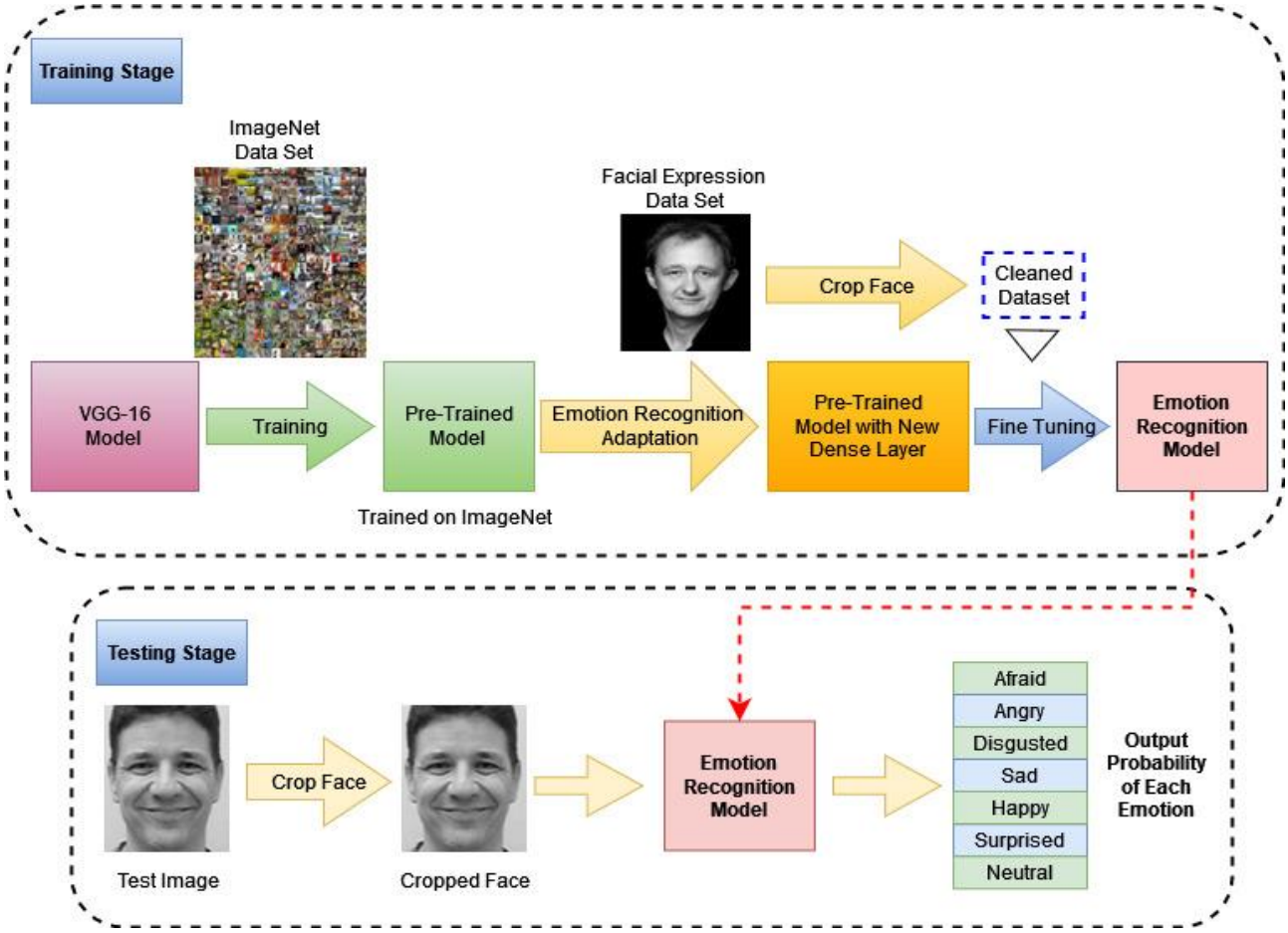


Figure 1. Proposed System Framework

Figure 1 illustrates the process of training and testing an emotion recognition model. The top section, labeled "Training Stage," indicates that the model starts with the ImageNet dataset and undergoes initial training using the VGG-16 model. Subsequently, the model incorporates the Facial Expression Dataset, where faces are cropped from images. The pretrained model then undergoes adaptation for emotion recognition and further fine-tuning before being refined and transformed into the Emotion Recognition Model. The bottom section, labeled "Testing Stage," depicts how test images go through the process of face cropping to produce cropped faces, which are then fed into the Emotion Recognition Model. The outcome is a probability distribution of emotions for seven categories: Angry, Disgusted, Fearful, Happy, Sad, Surprised, Neutral. This flowchart outlines the steps involved in creating an AI system capable of recognizing human emotions from facial expressions—a complex task that combines computer vision techniques and machine learning.

This proposed work focuses on facial expression recognition using two well-known convolutional neural network (CNN) architectures: ResNet50 and ResNet34. Facial expression recognition aims to detect and categorize emotions through facial cues, facilitating applications in emotion analysis, human-computer interaction, and affective computing.

a. **ResNet50 Architecture**: ResNet50 is a deep CNN architecture with 50 layers, notable for its residual connections that address the degradation problem encountered when training very deep neural networks. In this proposed work, ResNet50 will be used as the primary architecture for facial expression recognition [42].
b. **ResNet34 Architecture**: ResNet34 is a variant of the ResNet architecture with 34 layers. Although it has fewer layers than ResNet50, it still benefits from residual connections, enhancing performance and accuracy in various computer vision tasks. In this work, ResNet34 will be used as a comparative architecture to evaluate its performance in facial expression recognition [43].

The proposed system involves training and fine-tuning the ResNet50 and ResNet34 architectures using an appropriate dataset for facial expression recognition. This dataset will comprise facial images annotated with corresponding emotion labels. The system development process generally includes the following steps:

2.1 Data Collection

The first step involves gathering a comprehensive and labeled dataset of facial images. This dataset should cover diverse demographics and capture different individuals displaying these emotions in various contexts and lighting conditions. The primary objective is to ensure that the dataset is diverse and representative enough to train a robust facial expression recognition model effectively. By collecting a wide array of facial expressions, the model can learn to accurately identify

and classify emotions across different individuals and scenarios, thus enhancing its performance and generalization capabilities [44].

2.2 Data Preprocessing

In the data preprocessing phase, essential tasks are performed on the collected dataset to enhance its quality and prepare it for training. This includes face detection, which involves identifying and isolating faces within the images. Subsequently, alignment techniques are applied to ensure that all detected faces are aligned to a standardized orientation, facilitating consistent feature extraction. Additionally, normalization procedures are employed to adjust the images in terms of size, illumination, and color, thereby ensuring uniformity across the dataset. These preprocessing steps are crucial for improving the input data quality and optimizing the performance of the facial expression recognition model. The primary objective of data preprocessing is to enhance the quality and uniformity of the input data, ultimately leading to improved model performance and accuracy. By performing tasks such as face detection, alignment, and normalization, potential variations and inconsistencies within the dataset are minimized, enabling the model to effectively learn and extract meaningful features from the facial images [45].

2.3 Model Selection

The selection process involves choosing ResNet50 and ResNet34 as the neural network architectures for the facial expression classification task. These architectures are renowned for their effectiveness in feature extraction and image classification tasks, particularly in the field of computer vision. ResNet50 and ResNet34 are chosen due to their deep structure and residual connections, which enable them to capture intricate patterns and nuances in facial expressions. Additionally, these architectures have been extensively studied and validated in various research domains, demonstrating robust performance and accuracy. By selecting ResNet50 and ResNet34, the goal is to leverage the strengths of these proven architectures to develop a facial expression recognition system that can accurately classify emotions from facial images with high efficiency and reliability [46].

2.4 Transfer Learning Initialization

In the transfer learning initialization phase, pre-trained ResNet50 and ResNet34 models are employed, which have been previously trained on large-scale datasets. These models have already learned to extract meaningful features from a wide range of images, including diverse facial expressions, through their extensive training on extensive datasets. By leveraging these pre-trained models, the process of capturing relevant features from the input facial images is streamlined and made more efficient. This approach also reduces the need for extensive training data and time-consuming training procedures since the models have already acquired knowledge about various visual features present in facial expressions [47].

2.5 Model Fine-Tuning

The model fine-tuning phase involves initializing the weights of the pre-trained ResNet50 and ResNet34 models and adjusting them to better suit the specific facial expression recognition task. This process allows the models to adapt and refine their learned representations based on the nuances and intricacies of the facial expression dataset. By fine-tuning the pre-trained models, adjustments are made to optimize their performance for accurately classifying facial expressions. This adaptation ensures that the models can effectively capture the subtle variations in facial features indicative of different emotions, thereby improving their ability to make precise predictions. The objective of model fine-tuning is to improve the performance and accuracy of models in facial expression recognition by tailoring them to the unique characteristics of the target dataset. Through this iterative process, the models become more proficient at accurately identifying and classifying emotions from facial images, ultimately resulting in improved overall performance [48].

2.6 Dataset Splitting

The dataset splitting phase involves dividing the preprocessed dataset into distinct training and validation sets. This step is crucial for ensuring that the model can be effectively trained and evaluated. The training set comprises a significant portion of the data and is used to train the ResNet50 and ResNet34 models. Meanwhile, the validation set is used to monitor the models' performance during training and to tune hyperparameters. By evaluating the model on the validation set, adjustments can be made to optimize its performance and generalization capabilities. This division helps in preventing overfitting and ensures that the model can generalize well to unseen data. The primary objective of splitting the dataset is to utilize the training set for developing the model and the validation set for assessing its performance and fine-tuning hyperparameters. This approach allows for iterative improvements in the model's accuracy and robustness, ensuring that it performs well not only on the training data but also on new, unseen data [49].

2.7 Model Training

The model training phase involves training the ResNet50 and ResNet34 models using the training dataset. During this process, images from the training set are fed through the networks, allowing the models to learn and extract relevant features from the facial images. By minimizing the loss function, these optimization techniques help in refining the weights and biases of the models to enhance their accuracy and performance. Throughout the training process, the models continuously improve their ability to identify patterns and nuances in the facial expressions, thereby becoming more adept

at classifying different emotions accurately. The primary objective of model training is to enable the models to learn how to effectively extract features from the input images and make precise predictions. By using optimization techniques like SGD, the models' parameters are fine-tuned to minimize errors and improve their classification accuracy [50].

2.8 Model Validation

The model validation phase involves using the validation set to monitor the performance of the ResNet50 and ResNet34 models during training. By evaluating the models on the validation set, it is possible to gain insights into their accuracy and ability to generalize to new data. During this phase, hyperparameters such as learning rate, batch size, and the number of epochs is tuned to enhance the models' performance. Adjustments to these hyperparameters are made based on the validation performance, aiming to find the optimal configuration that balances training efficiency and accuracy. This iterative process helps in identifying and mitigating issues like overfitting. By continuously monitoring and adjusting hyperparameters based on the validation set performance, the models are fine-tuned to achieve better accuracy and generalization. This step is crucial for developing robust facial expression recognition models that maintain high performance across different datasets and real-world scenarios [51].

2.9 Feature Extraction and Prediction

During training, the ResNet50 and ResNet34 models extract features from the images using convolutional layers, apply non-linear transformations, and make predictions through fully connected layers with SoftMax activation. This process allows the models to learn hierarchical representations of facial features that are critical for accurate emotion classification. The primary objective is to enable the models to classify facial expressions accurately. By extracting and transforming relevant features, the models can differentiate between various emotions. Additionally, this process ensures that the models do not overfit and perform well on unseen data, maintaining high accuracy and generalization capabilities [52].

2.10 Model Evaluation

After training, the models are evaluated on a test set to assess their performance. This involves calculating evaluation metrics such as accuracy, precision, recall, and F1-score. These metrics provide a comprehensive understanding of how well the models can classify facial expressions and identify different emotions accurately. The primary objective is to measure the models' ability to correctly classify facial expressions and ensure their reliability. By evaluating these metrics, the robustness and effectiveness of the models are validated, confirming their readiness for deployment in real-world applications where accurate emotion recognition is essential [53].

2.11 Deployment and Application

Utilize the trained ResNet50 and ResNet34 models for real-world applications like affective computing, human-computer interaction, and emotion analysis to enhance practical functionalities. This involves feeding new, unseen facial images through the networks to obtain predicted emotion labels, allowing for real-time emotion recognition. The primary objective is to utilize the models' predictions to gain insights into individuals' emotional states. By applying this knowledge in various real-world scenarios, the models can enhance user experiences, improve interactions in human-computer interfaces, and contribute to fields requiring accurate emotion analysis.

The proposed system utilizes ResNet50 and ResNet34 architectures to harness the capabilities of deep learning and convolutional neural networks (CNNs) for precise facial expression recognition and classification. By comparing these two architectures, valuable insights can be gained regarding their effectiveness and suitability for facial expression recognition tasks.

ResNet34 and ResNet50 are intricate convolutional neural network architectures composed of numerous layers. Providing a detailed account of the complete mathematical expressions for these architectures, along with the facial dataset, would be exceedingly lengthy and difficult to cover thoroughly in a text-based format. Nonetheless, we can offer a high-level overview of the architecture and the essential mathematical operations involved. Here is a simplified explanation:

ResNet34: ResNet34 is composed of 34 layers, including residual blocks. The mathematical representation of a residual block in ResNet34 can be expressed as [54]:

$$y = F(x) + x \tag{1}$$

In this context, x denotes the input feature map, $F(x)$ signifies the non-linear transformations executed by the residual block, and y represents the output feature map.

ResNet50: ResNet50 features a more intricate architecture with 50 layers. Like ResNet34, it employs residual blocks. The mathematical representation of a residual block in ResNet50 can be expressed as [55]:

$$y = F(x) + W_S * x \tag{2}$$

In this context, x is the input feature map, $F(x)$ denotes the non-linear transformations carried out by the residual block, y is the output feature map, and Ws is a learnable weight matrix.

3. RESULTS AND DISCUSSION

3.1 Experimental Results

This section presents a comprehensive experimental assessment of the model's performance across diverse facial expression recognition databases. Initially, a concise introduction to the databases employed in the study is presented. Following this, the models' performance on four distinct databases is detailed. Subsequently, a comparison with recent research findings in the domain is conducted. Lastly, a visualization method is utilized to emphasize the significant regions identified by the trained model.

The Facial Expression Recognition 2013 (FER2013) database, unveiled at the ICML 2013 Challenges in Representation Learning, comprises 35,887 images predominantly captured in natural settings and sized at 48x48 pixels. Initially, the training set consisted of 28,709 images, while both the validation and test sets contained 3,589 images each. Utilizing the Google Image Search API facilitated dataset compilation, with faces automatically added during the process. All facial expressions, including neutral ones, are categorized into one of the six primary facial moods. FER2013 stands out from other datasets due to its inclusion of more diverse images featuring various characteristics, such as facial occlusions (often due to a hand), partial faces, low-contrast images, and individuals wearing glasses. Figure 2 offers a glimpse of four typical images from the FER2013 dataset, showcasing its diverse and challenging nature.

Figure 2. Dataset

Confusion Matrix - The confusion matrix depicted in Figure 2, derived from applying the proposed model to the FER dataset's validation set, reveals a trend where the model tends to exhibit more errors with classes characterized by fewer instances, notably disgust and fear.

Model Visualization - This study presents a straightforward method for identifying key facial regions crucial for recognizing a range of emotions. The method begins by systematically occluding an N-by-N-pixel square. Subsequently, the trained model predicts the emotion based on the occluded image. Should occlusion of a particular area result in an erroneous prediction, that region is deemed significant for recognizing that emotion. Conversely, if the prediction remains unaltered, the region is considered less crucial. This process is iterated with numerous sliding windows of size N-by-N, each shifted across the image. By analyzing the impact of occlusion on predictions across various regions, the model identifies key areas pivotal for accurate emotion recognition.

Figure 3. Frame work of Proposed System

Figure 3 shows a graphical user interface (GUI) that is part of a software application associated with machine learning or deep learning models. Each button likely corresponds to a different neural network architecture or function within the application. This interface is used to select or compare these models, possibly for tasks such as image recognition or classification.

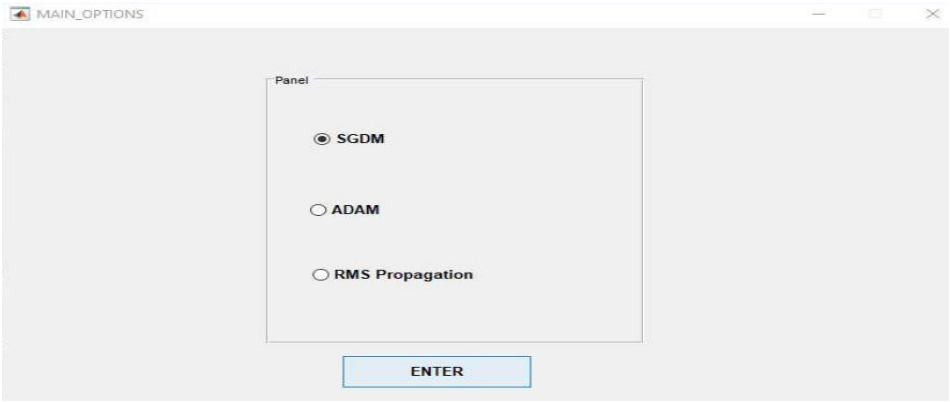
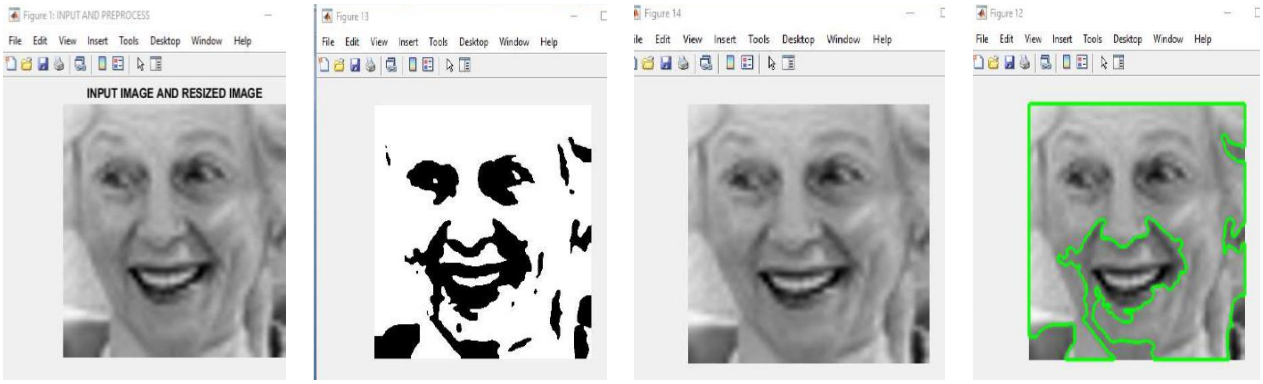


Figure 4. Optimization Network

Figure 4 displays a panel interface featuring three buttons, each offering a different optimization algorithm option. These algorithms can be employed in machine learning to enhance performance, providing users with the ability to select the most suitable method for their specific needs.



Input Image Binary Image Grey Image Feature Image

Figure 5. Image Pre-processing and Segmentation Result

Figure 5 shows a series of images demonstrating the transformation of an input image through multiple processing stages to extract pertinent features for analysis. These steps are crucial for applications in computer vision and machine learning. Initially, the image is resized, then converted to a binary format, followed by a grayscale version. The final step highlights the extraction of specific features. Such detailed preprocessing and segmentation processes are essential for accurately analyzing and interpreting image data in various advanced technological fields.

3.2 Performance Analysis

In assessing the performance of facial expression recognition systems, a range of evaluation metrics can be utilized, encompassing true positives (TP), true negatives (TN), false positives (FP), false negatives (FN), accuracy, precision, recall (or sensitivity), and specificity. These metrics offer a comprehensive overview of the system's effectiveness in correctly identifying positive and negative instances, as well as its overall accuracy, precision, and ability to detect true positives while minimizing false positives and negatives.

True Positives (TP): The count of correctly predicted positive samples, indicating facial expressions that were accurately classified as positive (correctly recognized emotions).

True Negatives (TN): The count of correctly predicted negative samples, signifying facial expressions that were correctly classified as negative (correctly recognized as non-target emotions).

False Positives (FP): The count of incorrectly predicted positive samples, denoting facial expressions that were classified as positive but should have been classified as negative (misclassified as the wrong emotion).

False Negatives (FN): The count of incorrectly predicted negative samples, representing facial expressions that were classified as negative but should have been classified as positive (missed detection of the target emotion).

Before delving into these evaluation metrics, it's essential to understand how they provide insights into the performance of facial expression recognition models. Accuracy, precision, recall, and specificity offer valuable perspectives on the model's ability to classify facial expressions accurately. Let us explore each metric and its significance in evaluating model performance.

Accuracy, defined as the proportion of accurately classified samples relative to the total sample size, is calculated as follows:

$$\frac{TP+TN}{TP+TN+FP+FN} \tag{3}$$

It offers a comprehensive assessment of the model's performance in facial expression recognition.

Precision, synonymous with positive predictive value, evaluates the fraction of correctly identified positive samples out of all samples classified as positive, and is computed as follows:

$$\frac{TP}{TP+FP} \tag{4}$$

It reflects the model's capability to minimize false positive predictions.

Recall, also known as the true positive rate or sensitivity, measures the percentage of correctly identified positive samples out of all actual positive samples, and is determined by the following formula:

$$\frac{TP}{TP+FN} \tag{5}$$

It signifies the model's effectiveness in accurately detecting positive samples.

Specificity, sometimes referred to as the "true negative rate," quantifies the number of correctly predicted negative samples among all the actual negative samples. Here's the method to determine specificity:

$$\frac{TN}{TN+FP} \tag{6}$$

It demonstrates the model's capacity to accurately identify non-target emotions

The test results of the proposed system can be seen in the following table:

Table 1. Performance of the Propose System

Panel	Deep Learning Techniques	Accuracy (%)	Precision (%)	Recall (%)	Specificity (%)
SGDM	ResNet34 and ResNet50	98.19	97.26	97.26	97.15
ADAM		98.23	96.25	96.25	98.23
RMS Propagation		98.24	98.21	98.21	96.23

Table 1 illustrates the performance of the proposed system employing different deep learning techniques, namely SGDM, ADAM, and RMS Propagation, in conjunction with ResNet34 and ResNet50 architectures. Each technique underwent evaluation based on key metrics including accuracy, precision, recall, and specificity. SGDM attained an accuracy of 98.19%, with both precision and recall registering at 97.26%, and specificity at 97.15%. ADAM exhibited a slightly higher accuracy of 98.23%, albeit with lower precision and recall at 96.25%. However, ADAM demonstrated a higher specificity of 98.23%. Meanwhile, RMS Propagation recorded the highest accuracy at 98.24%, with precision and recall mirroring each other at 98.21%. Its specificity, though slightly lower, remained notable at 96.23%.

The proposed system, employing ResNet50 and ResNet34, achieved a remarkable accuracy of 98.19%, outperforming previous studies. Mohammed et al reported an accuracy of 68% using a CNN architecture [56], while Xiaoqing et al achieved 65.3% with unsupervised domain adaptation [57]. Additionally, methods such as, VGG+SVM [58], GoogleNet [59], and FER on SoC [60] yielded accuracies ranging from 65.2% to 66.31%. Notably, Aff-Wild2 with a VGG backbone achieved the highest accuracy of 75% [61]. The significant improvement in accuracy demonstrated by the proposed system underscores the effectiveness of utilizing ResNet50 and ResNet34 architectures in facial expression recognition tasks compared to other deep learning techniques and methodologies utilized in prior studies.

4. CONCLUSION

This research presents a novel method for facial expression recognition employing an attentional convolutional network, underscoring the significance of concentrating on facial areas to achieve precise emotion detection. By leveraging this approach, hybrid neural networks, combining ResNet50 and ResNet34 architectures, demonstrate promising advancements in the field. Through extensive experimentation across four widely used facial expression databases, significant progress is observed. The incorporation of visualization techniques highlights salient facial features crucial for distinguishing between different emotions. The hybrid ResNet50 and ResNet34 architecture, facilitated by transfer learning and fine-tuning, effectively captures meaningful facial features, enhancing emotion recognition accuracy and generalization. The deeper architecture of ResNet50 enables learning complex patterns, while the lightweight ResNet34 offers computational efficiency. This balance between performance and efficiency makes the hybrid approach versatile for facial expression classification. Comprehensive evaluation using metrics like accuracy, precision, recall, and F1-score underscores the reliability and robustness of the hybrid ResNet50 and ResNet34 architecture. The potential applications span affective computing, human-computer interaction, and emotion analysis, enabling more accurate interpretation of human emotions from facial images and the development of intelligent systems. However, the hybrid approach presents challenges, including computational demands and the need for diverse and representative training data. Addressing these limitations will be crucial for maximizing the potential of the hybrid ResNet50 and ResNet34 architecture in real-world applications, ensuring its effectiveness across various contexts and demographic groups.

REFERENCES

[1] L. Sels, H. T. Reis, A. K. Randall, and L. Verhofstadt, "Emotion Dynamics in Intimate Relationships: The Roles of Interdependence and Perceived Partner Responsiveness," in Affect Dynamics, Cham: Springer International Publishing, 2021, pp. 155–179. doi: 10.1007/978-3-030-82965-0_8.

[2] S. C. Leong, Y. M. Tang, C. H. Lai, and C. K. M. Lee, "Facial expression and body gesture emotion recognition: A systematic review on the use of visual data in affective computing," Comput. Sci. Rev., vol. 48, p. 100545, May 2023, doi: 10.1016/j.cosrev.2023.100545.

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

[4] E. G. Krumhuber, L. I. Skora, H. C. H. Hill, and K. Lander, "The role of facial movements in emotion recognition," Nat. Rev. Psychol., vol. 2, no. 5, pp. 283–296, Mar. 2023, doi: 10.1038/s44159-023-00172-1.

[5] M. A. Solis-Arrazola, R. E. Sanchez-Yañez, C. H. Garcia-Capulin, and H. Rostro-Gonzalez, "Enhancing image-based facial expression recognition through muscle activation-based facial feature extraction," Comput. Vis. Image Underst., vol. 240, p. 103927, Mar. 2024, doi: 10.1016/j.cviu.2024.103927.

[6] H. Pérez-Espinosa, R. Zatarain-Cabada, and M. L. Barrón-Estrada, "Emotion recognition: from speech and facial expressions," in Biosignal Processing and Classification Using Computational Learning and Intelligence, Elsevier, 2022, pp. 307–326. doi: 10.1016/B978-0-12-820125-1.00028-2.

[7] A. R. Dores, F. Barbosa, C. Queirós, I. P. Carvalho, and M. D. Griffiths, "Recognizing Emotions through Facial Expressions: A Largescale Experimental Study," Int. J. Environ. Res. Public Health, vol. 17, no. 20, p. 7420, Oct. 2020, doi: 10.3390/ijerph17207420.

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

[9] 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, 2023, doi: 10.1016/j.matpr.2021.07.297.

[10] H. Ghazouani, "Challenges and Emerging Trends for Machine Reading of the Mind from Facial Expressions," SN Comput. Sci., vol. 5, no. 1, p. 103, Dec. 2023, doi: 10.1007/s42979-023-02447-z.

[11] Nidhi and B. Verma, "From methods to datasets: a detailed study on facial emotion recognition," Appl. Intell., vol. 53, no. 24, pp. 30219–30249, Dec. 2023, doi: 10.1007/s10489-023-05052-y.

[12] M. Sajjad et al., "A comprehensive survey on deep facial expression recognition: challenges, applications, and future guidelines," Alexandria Eng. J., vol. 68, pp. 817–840, Apr. 2023, doi: 10.1016/j.aej.2023.01.017.

[13] S. H. Syed and V. Muralidharan, "Feature extraction using Discrete Wavelet Transform for fault classification of planetary gearbox – A comparative study," Appl. Acoust., vol. 188, p. 108572, Jan. 2022, doi: 10.1016/j.apacoust.2021.108572.

[14] M. Zuhaib et al., "Faults Feature Extraction Using Discrete Wavelet Transform and Artificial Neural Network for Induction Motor Availability Monitoring—Internet of Things Enabled Environment," Energies, vol. 15, no. 21, p. 7888, Oct. 2022, doi: 10.3390/en15217888.

[15] A. Topic and M. Russo, "Emotion recognition based on EEG feature maps through deep learning network," Eng. Sci. Technol. an Int. J., vol. 24, no. 6, pp. 1442–1454, Dec. 2021, doi: 10.1016/j.jestch.2021.03.012.

[16] R. Yuvaraj, P. Thagavel, J. Thomas, J. Fogarty, and F. Ali, "Comprehensive Analysis of Feature Extraction Methods for Emotion Recognition from Multichannel EEG Recordings," Sensors, vol. 23, no. 2, p. 915, Jan. 2023, doi: 10.3390/s23020915.

[17] J. Zhang, Z. Yin, P. Chen, and S. Nichele, "Emotion recognition using multi-modal data and machine learning techniques: A tutorial and review," Inf. Fusion, vol. 59, pp. 103–126, Jul. 2020, doi: 10.1016/j.inffus.2020.01.011.

[18] A. Khan, A. Sohail, U. Zahoora, and A. S. Qureshi, "A survey of the recent architectures of deep convolutional neural networks," Artif. Intell. Rev., vol. 53, no. 8, pp. 5455–5516, Dec. 2020, doi: 10.1007/s10462-020-09825-6.

[19] I. Lobato, T. Friedrich, and S. Van Aert, "Deep convolutional neural networks to restore single-shot electron microscopy images," npj Comput. Mater., vol. 10, no. 1, p. 10, Jan. 2024, doi: 10.1038/s41524-023-01188-0.

[20] X. Zhao, L. Wang, Y. Zhang, X. Han, M. Deveci, and M. Parmar, "A review of convolutional neural networks in computer vision," Artif. Intell. Rev., vol. 57, no. 4, p. 99, Mar. 2024, doi: 10.1007/s10462-024-10721-6.

[21] H. Li, W. Wang, M. Wang, L. Li, and V. Vimlund, "A review of deep learning methods for pixel-level crack detection," J. Traffic Transp. Eng. (English Ed., vol. 9, no. 6, pp. 945–968, Dec. 2022, doi: 10.1016/j.jtte.2022.11.003.

[22] S. Sony, K. Dunphy, A. Sadhu, and M. Capretz, "A systematic review of convolutional neural network-based structural condition assessment techniques," Eng. Struct., vol. 226, p. 111347, Jan. 2021, doi: 10.1016/j.engstruct.2020.111347.

[23] D. Zhao, Y. Qian, J. Liu, and M. Yang, "The facial expression recognition technology under image processing and neural network," J. Supercomput., vol. 78, no. 4, pp. 4681–4708, Mar. 2022, doi: 10.1007/s11227-021-04058-y.

[24] B. Ko, "A Brief Review of Facial Emotion Recognition Based on Visual Information," Sensors, vol. 18, no. 2, p. 401, Jan. 2018, doi: 10.3390/s18020401.

[25] S. Park et al., "Differences in Facial Expressions between Spontaneous and Posed Smiles: Automated Method by Action Units and Three-Dimensional Facial Landmarks," Sensors, vol. 20, no. 4, p. 1199, Feb. 2020, doi: 10.3390/s20041199.

[26] M. T. Mustapha, I. Ozsahin, and D. U. Ozsahin, "Convolution neural network and deep learning," in Artificial Intelligence and Image Processing in Medical Imaging, Elsevier, 2024, pp. 21–50. doi: 10.1016/B978-0-323-95462-4.00002-9.

[27] S. R. Shah, S. Qadri, H. Bibi, S. M. W. Shah, M. I. Sharif, and F. Marinello, "Comparing Inception V3, VGG 16, VGG 19, CNN, and ResNet 50: A Case Study on Early Detection of a Rice Disease," Agronomy, vol. 13, no. 6, p. 1633, Jun. 2023, doi: 10.3390/agronomy13061633.

[28] K. K. Bressem, L. C. Adams, C. Erxleben, B. Hamm, S. M. Niehues, and J. L. Vahldiek, "Comparing different deep learning architectures for classification of chest radiographs," Sci. Rep., vol. 10, no. 1, p. 13590, Aug. 2020, doi: 10.1038/s41598-020-70479-z.

[29] A. Victor Ikechukwu, S. Murali, R. Deepu, and R. C. Shivamurthy, "ResNet-50 vs VGG-19 vs training from scratch: A comparative analysis of the segmentation and classification of Pneumonia from chest X-ray images," Glob. Transitions Proc., vol. 2, no. 2, pp. 375–381, Nov. 2021, doi: 10.1016/j.gltp.2021.08.027.

[30] S. K. Khare, V. Blanes-Vidal, E. S. Nadimi, and U. R. Acharya, "Emotion recognition and artificial intelligence: A systematic review (2014–2023) and research recommendations," Inf. Fusion, vol. 102, p. 102019, Feb. 2024, doi: 10.1016/j.inffus.2023.102019.
[31] E. A. Clark et al., "The Facial Action Coding System for Characterization of Human Affective Response to Consumer Product-Based Stimuli: A Systematic Review," Front. Psychol., vol. 11, May 2020, doi: 10.3389/fpsyg.2020.00920.
[32] R. Yelchuri, J. K. Dash, P. Singh, A. Mahapatro, and S. Panigrahi, "Exploiting deep and hand-crafted features for texture image retrieval using class membership," Pattern Recognit. Lett., vol. 160, pp. 163–171, Aug. 2022, doi: 10.1016/j.patrec.2022.06.017.
[33] L. Alzubaidi et al., "Review of deep learning: concepts, CNN architectures, challenges, applications, future directions," J. Big Data, vol. 8, no. 1, p. 53, Mar. 2021, doi: 10.1186/s40537-021-00444-8.
[34] J. Chai, H. Zeng, A. Li, and E. W. T. Ngai, "Deep learning in computer vision: A critical review of emerging techniques and application scenarios," Mach. Learn. with Appl., vol. 6, p. 100134, Dec. 2021, doi: 10.1016/j.mlwa.2021.100134.
[35] I. H. Sarker, "Deep Learning: A Comprehensive Overview on Techniques, Taxonomy, Applications and Research Directions," SN Comput. Sci., vol. 2, no. 6, p. 420, Nov. 2021, doi: 10.1007/s42979-021-00815-1.
[36] X. Zhao, X. Shi, and S. Zhang, "Facial Expression Recognition via Deep Learning," IETE Tech. Rev., vol. 32, no. 5, pp. 347–355, Sep. 2015, doi: 10.1080/02564602.2015.1017542.
[37] J. Li, K. Jin, D. Zhou, N. Kubota, and Z. Ju, "Attention mechanism-based CNN for facial expression recognition," Neurocomputing, vol. 411, pp. 340–350, Oct. 2020, doi: 10.1016/j.neucom.2020.06.014.
[38] R. Bendjillali, M. Beladgham, K. Merit, and A. Taleb-Ahmed, "Improved Facial Expression Recognition Based on DWT Feature for Deep CNN," Electronics, vol. 8, no. 3, p. 324, Mar. 2019, doi: 10.3390/electronics8030324.
[39] Q. T. Ngoc, S. Lee, and B. C. Song, "Facial Landmark-Based Emotion Recognition via Directed Graph Neural Network," Electronics, vol. 9, no. 5, p. 764, May 2020, doi: 10.3390/electronics9050764.
[40] M. Reyad, A. M. Sarhan, and M. Arafa, "A modified Adam algorithm for deep neural network optimization," Neural Comput. Appl., vol. 35, no. 23, pp. 17095–17112, Aug. 2023, doi: 10.1007/s00521-023-08568-z.
[41] E. Hassan, M. Y. Shams, N. A. Hikal, and S. Elmougy, "The effect of choosing optimizer algorithms to improve computer vision tasks: a comparative study," Multimed. Tools Appl., vol. 82, no. 11, pp. 16591–16633, May 2023, doi: 10.1007/s11042-022-13820-0.
[42] R. I. Bendjillali, M. Beladgham, K. Merit, and A. Taleb-Ahmed, "Illumination-robust face recognition based on deep convolutional neural networks architectures," Indones. J. Electr. Eng. Comput. Sci., vol. 18, no. 2, p. 1015, May 2020, doi: 10.11591/ijeecs.v18.i2.pp1015-1027.
[43] N. M. Dipu, S. Alam Shohan, and K. M. . Salam, "Ocular Disease Detection Using Advanced Neural Network Based Classification Algorithms," ASIAN J. Converg. Technol., vol. 7, no. 2, pp. 91–99, Aug. 2021, doi: 10.33130/AJCT.2021v07i02.019.
[44] R. R*, S. Lal P, A. M. Philip, and V. V, "A Compact Deep Learning Model for Robust Facial Expression Recognition.," Int. J. Eng. Adv. Technol., vol. 8, no. 6, pp. 2956–2960, Aug. 2019, doi: 10.35940/ijeat.F8724.08861.
[45] Y. Gong, G. Liu, Y. Xue, R. Li, and L. Meng, "A survey on dataset quality in machine learning," Inf. Softw. Technol., vol. 162, p. 107268, Oct. 2023, doi: 10.1016/j.infsof.2023.107268.
[46] 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," 2023, pp. 281–300. doi: 10.1007/978-981-99-1624-5_21.
[47] A. Soliman, S. Shaheen, and M. Hadhoud, "Leveraging pre-trained language models for code generation," Complex Intell. Syst., vol. 10, no. 3, pp. 3955–3980, Jun. 2024, doi: 10.1007/s40747-024-01373-8.
[48] N. Mehendale, "Facial emotion recognition using convolutional neural networks (FERC)," SN Appl. Sci., vol. 2, no. 3, p. 446, Mar. 2020, doi: 10.1007/s42452-020-2234-1.
[49] V. R. Joseph and A. Vakayil, "SPlit: An Optimal Method for Data Splitting," Technometrics, vol. 64, no. 2, pp. 166–176, Apr. 2022, doi: 10.1080/00401706.2021.1921037.
[50] S. F. Ahmed et al., "Deep learning modelling techniques: current progress, applications, advantages, and challenges," Artif. Intell. Rev., vol. 56, no. 11, pp. 13521–13617, Nov. 2023, doi: 10.1007/s10462-023-10466-8.
[51] M. Wojciuk, Z. Swiderska-Chadaj, K. Siwek, and A. Gertych, "Improving classification accuracy of fine-tuned CNN models: Impact of hyperparameter optimization," Heliyon, vol. 10, no. 5, p. e26586, Mar. 2024, doi: 10.1016/j.heliyon.2024.e26586.
[52] S. Rajan, P. Chenniappan, S. Devaraj, and N. Madian, "Facial expression recognition techniques: a comprehensive survey," IET Image Process., vol. 13, no. 7, pp. 1031–1040, May 2019, doi: 10.1049/iet-ipr.2018.6647.
[53] M. J. A. Dujaili, "Survey on facial expressions recognition: databases, features and classification schemes," Multimed. Tools Appl., vol. 83, no. 3, pp. 7457–7478, Jan. 2024, doi: 10.1007/s11042-023-15139-w.
[54] M. Gao, D. Qi, H. Mu, and J. Chen, "A Transfer Residual Neural Network Based on ResNet-34 for Detection of Wood Knot Defects," Forests, vol. 12, no. 2, p. 212, Feb. 2021, doi: 10.3390/f12020212.
[55] J. Yang et al., "Prediction of HER2-positive breast cancer recurrence and metastasis risk from histopathological images and clinical information via multimodal deep learning," Comput. Struct. Biotechnol. J., vol. 20, pp. 333–342, 2022, doi: 10.1016/j.csbj.2021.12.028.
[56] M. Aamir et al., "A deep learning approach for brain tumor classification using MRI images," Comput. Electr. Eng., vol. 101, p. 108105, Jul. 2022, doi: 10.1016/j.compeleceng.2022.108105.
[57] X. Wang, X. Wang, and Y. Ni, "Unsupervised Domain Adaptation for Facial Expression Recognition Using Generative Adversarial Networks," Comput. Intell. Neurosci., vol. 2018, pp. 1–10, Jul. 2018, doi: 10.1155/2018/7208794.
[58] M.-I. Georgescu, R. T. Ionescu, and M. Popescu, "Local Learning With Deep and Handcrafted Features for Facial Expression Recognition," IEEE Access, vol. 7, pp. 64827–64836, 2019, doi: 10.1109/ACCESS.2019.2917266.
[59] P. Giannopoulos, I. Perikos, and I. Hatzilygeroudis, "Deep Learning Approaches for Facial Emotion Recognition: A Case Study on FER-2013," 2018, pp. 1–16. doi: 10.1007/978-3-319-66790-4_1.
[60] P. T. Vinh and T. Quang Vinh, "Facial Expression Recognition System on SoC FPGA," in 2019 International Symposium on Electrical and Electronics Engineering (ISEE), Oct. 2019, pp. 1–4. doi: 10.1109/ISEE2.2019.8921140.
[61] D. Kollias and S. Zafeiriou, "Expression, affect, action unit recognition: Aff-wild2, multi-task learning and arcface," 30th Br. Mach. Vis. Conf. 2019, BMVC 2019, no. October, 2020.