International Journal of Computing and Digital Systems
ISSN (2210-142X)
Int. J. Com. Dig. Sys. 15, No.1 (May-24)

http://dx.doi.org/10.12785/ijcds/1501117

Performance Evaluation of Deep Learning Models for Face Expression Recognition

Raed Ibrahim Khaleel Almsari[1], Abbas Hussein Miry[2] and Tariq M. Salman [3]

[1,2,3]*Electrical Engineering Department Al-mustansiriyah University Baghdad-Iraq*

Received 15 Oct. 2023, Revised 19 Mar. 2024, Accepted 20 Mar. 2024, Published 1 May. 2024

Abstract: Facial expression recognition presents a significant challenge in computer vision, crucial for various applications like human-computer interaction and emotion analysis. Despite its importance, accurately discerning emotions from facial images remains complex due to factors such as lighting variations, pose differences, and subtle expression nuances. In this study, we aim to comprehensively evaluate five deep learning models - CNN, VGG16, Inception V3, MobileNet V2, and DenseNet121 - utilizing the CK+ dataset. Our research seeks to clarify the objectives and contributions early, emphasizing the significance of facial expression recognition. We provide an overview of the paper's structure to guide the reader through the logical progression of ideas. The background and related work section reviews existing literature, highlighting recent advancements and identifying research gaps. The methodology details dataset characteristics, preprocessing steps, and model architectures, followed by the experimental results section presenting performance metrics and comparisons. The discussion interprets results, analyzing model strengths and weaknesses while considering practical implications and future research directions. Finally, the conclusion summarizes key findings and emphasizes the study's significance, suggesting avenues for further exploration. Throughout the paper, clarity, readability, and grammatical accuracy are maintained, supported by visual aids like tables or diagrams where necessary to enhance comprehension. .

Keywords: Face Expression, Emotion Recognition, Deep Learning, Transfer Learning.

1. Introduction

Facial emotion recognition (FER) is a pivotal domain within artificial intelligence and computer vision, offering versatile applications in video analytics, human-computer interactions, and more. This field traces its roots back to the seminal work of Ekman and Friesen in 1978, highlighting the significance of human facial expressions in nonverbal communication. FER aims to create algorithms capable of categorising facial expressions into emotional states, with potential applications spanning healthcare, security, education, and beyond [1].

FER's practical applications encompass smart living, healthcare, human-computer and human-robot interactions, and augmented reality. Deep neural networks (DNNs), particularly convolutional neural networks (CNNs), have played a crucial role in advancing FER due to their ability to combine feature extraction and recognition [2]. However, CNN-based FER faces several challenges, including subtle differences between emotional expressions, intra-class variation, limited inter-class distinction, and variations caused by facial position changes and occlusions [3].

Despite impressive progress in controlled environments, FER encounters challenges in real-world scenarios. Deep CNN architectures offer promise but require extensive training and resources. Pre-training using models like EfficientNet, DenseNet, Inception-v3, Resnet-50, and VGG-16 has shown potential but demands substantial datasets and computational power. .

Facial expressions serve as external manifestations of ongoing or evolving mental processes [4]. They act as reliable indicators of these internal changes, primarily because a specific set of facial muscles consistently accompanies each expression. Deciphering a person's emotions becomes relatively straightforward when observing these expressions. Facial cues can reveal whether an individual is unaware of something or is attempting to conceal it. Additionally, the context in which an expression occurs can significantly influence its interpretation. Human beings possess an innate ability to react to stimuli without formal instruction. In essence, our minds naturally strive to draw conclusions about a person's actions by connecting their expressions to specific meanings [5].

It is important to recognize that the meaning behind facial expressions may not always be accurately interpreted. Consider this scenario: a husband acts inappropriately at a social event, causing his wife to become angry. Despite her

E-mail address: raed@uomustansiriyah.edu.iq, abbasmiry83@uomustansiriyah.edu.iq, tjibori@gmail.com

anger, the wife tries to conceal her true emotions by smiling during the event, which could lead others to assume that she is happy. This incorrect assumption highlights the complexity and multi-dimensional nature of facial expressions [4]. A single facial expression can convey a combination of emotions; for example, a person might simultaneously feel sadness and shock. Human faces can evoke multiple emotions simultaneously. Unlike words, which can sometimes be used to deceive, facial expressions offer a trustworthy means of discerning a person's true intentions [6]. Facial expressions are the primary avenue through which emotions find expression, and at times, deliberate facial expressions are employed to convey specific information.

This paper explores the evolution of FER techniques, focusing on emotion extraction from facial expressions, leveraging CNN capabilities, and addressing training challenges. By delving into these complexities, we aim to discover innovative approaches for achieving higher accuracy and efficacy in facial emotion recognition.

In this paper, we delve into the evolution of Facial Emotion Recognition (FER) techniques, emphasizing emotion extraction from facial expressions, leveraging the capabilities of Convolutional Neural Networks (CNNs), and addressing training challenges. We begin by highlighting the significance of FER in artificial intelligence and computer vision, tracing its origins to the seminal work of Ekman and Friesen in 1978. We underscore the practical applications of FER across various domains such as video analytics, human-computer interactions, healthcare, security, and education.

Facial expressions are reliable indicators of internal mental processes, and deep CNN architectures have shown promise in advancing FER. However, CNN-based approaches face several challenges, including subtle differences between emotional expressions, intra-class variation, limited inter-class distinction, and variations caused by facial position changes and occlusions. Extensive training and resources are necessary for deep CNN architectures, requiring pre-training using models like EfficientNet, DenseNet, Inception-v3, Resnet-50, and VGG-16. This, in turn, requires substantial datasets and computational power. Facial expressions are complex and can be easily misinterpreted. We illustrate this complexity with a hypothetical scenario involving the misinterpretation of facial expressions in social interactions. Facial expressions can convey a combination of emotions simultaneously, unlike words. They offer a trustworthy means of discerning a person's true intentions, underscoring their importance in communication and interaction. .

2. Facial Emotion Recognition

Facial Expression Recognition (FER) stands as a prominent field within computer vision research, with its primary goal being the identification of six fundamental facial expressions: anger, disgust, fear, happiness, sadness, and surprise, from images. The typical FER research workflow consists of a two-step process: initially, expressive features are extracted from provided images or videos, followed by the training of classifiers to distinguish between different expressions [7]. Traditional approaches [8] depend on manually crafted features to represent emotional images, while recent years have witnessed the rise of Deep Learning techniques as effective solutions for the FER challenge [9]. Deep Learning models often exhibit superior performance due to their capacity to autonomously learn features and undergo end-to-end training.

However, FER remains a continuous challenge. On one hand, expressive behaviors tend to manifest in specific facial regions, such as the areas around the eyes or mouth. Consequently, features extracted from unrelated regions may introduce redundancy, potentially degrading overall performance [7]. On the other hand, factors like pose variations and subjects' appearances can compromise the quality of extracted features, making them less effective in classifying different expressions. This dual challenge highlights the complexities involved in achieving accurate and robust Facial Expression Recognition [10]

The core essence of the fundamental FER process, as depicted in Figure (1), has remained consistent for over a decade, despite the emergence of new deep learning techniques, algorithms, and FER datasets. This journey begins with a fundamental prerequisite shared by all FER methods: the availability of a high-quality, diverse, and balanced dataset [11]. It is from such a dataset that the best results can be achieved by implementing unique strategies. However, the effectiveness of these strategies relies on data preprocessing, a crucial step in eliminating unnecessary noise. Following preprocessing, a classifier is trained using deep learning techniques, and the resulting model is evaluated using performance metrics. These metrics play a pivotal role in determining whether the classifier's performance meets the desired standards [11]. In conclusion Facial Expression Recognition focuses on identifying and categorizing specific facial movements, Facial Emotion Recognition goes further by attempting to understand the underlying emotional states conveyed by those expressions, considering both the visible facial cues and the broader emotional context.

3. Deep and Transfer Learning

Deep learning is a subset of machine learning that focuses on training artificial neural networks with multiple layers, known as deep neural networks. These networks are designed to automatically learn and extract hierarchical features from the input data, making them particularly suited for tasks involving complex data like images, audio, and text. Key characteristics of deep learning include deep neural networks (DNNs) consisting of multiple interconnected layers, backpropagation for training, deep feature learning, and reliance on big data and GPU acceleration [12].

Transfer learning, on the other hand, is a machine learning technique that leverages pre-trained models, typically

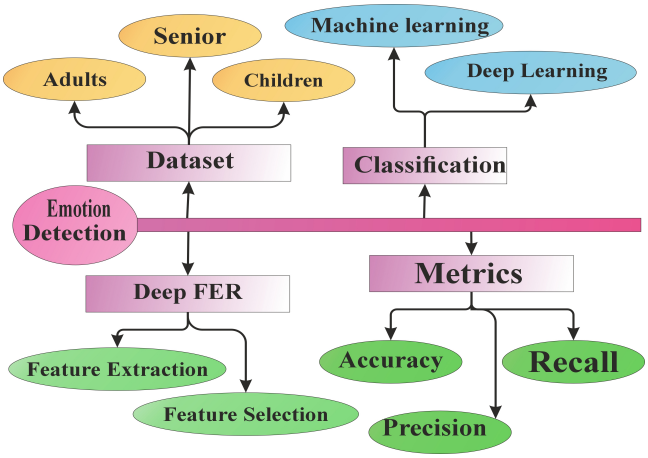

Figure 1. Overview of trends within facial emotion recognition

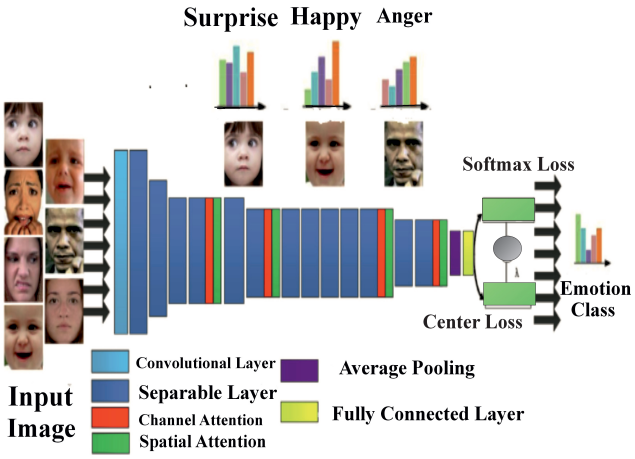

Figure 2. Use of Mobile Net V2 In emotion classification [16]

trained on large datasets for specific tasks, to adapt or fine-tune them for different but related tasks. It allows models to transfer knowledge gained from a source task to improve performance on a target task[13]. This is achieved through fine-tuning pre-trained models, and adjusting their parameters based on domain-specific data. Transfer learning is particularly useful when data for the target task is limited or when the source and target domains are related, as it can significantly enhance model performance in such scenarios [14].

In practice, deep learning and transfer learning often complement each other. Transfer learning enables the utilization of pre-trained deep models with rich feature representations, adapting them to specific tasks and reducing the need for extensive data and computational resources. This combination has played a pivotal role in advancing artificial intelligence applications across various domains, including computer vision, natural language processing, and speech recognition [15]. Figure (2) shows a sample of pre trained Model MobileNet V2 in Facial Expression classification.

Transfer learning offers the advantage of faster training, lower data requirements, and improved generalization by leveraging pre-trained models and their learned features from large datasets. This approach is resource-efficient, robust against overfitting, and often achieves state-of-the-art performance in various computer vision tasks. On the other hand, training deep CNNs from scratch requires more data, computational resources, and time, but provides finer control over model architecture. The choice between transfer learning and deep learning CNN depends on project-specific constraints, available resources, and the need for rapid prototyping or fine-tuned customization.

4. Related Works

Several related works in the field of Facial Expression Recognition (FER) have explored various techniques and datasets to achieve accurate emotion recognition. These studies have addressed the significance of FER for Human-Computer Interaction (HCI) [17] and have compared pre-trained Convolutional Neural Networks (CNNs) with transfer learning using Support Vector Machines (SVMs). They assessed performance on databases like CK+ and NVIE, achieving high accuracy.

While CNNs demonstrated effective integration with HCI, their data-hungry training and some handcrafted approach limitations were noted. Other studies have employed different methodologies, such as Local Fisher Discriminant Analysis (LFDA) and k-Nearest Neighbors (k-NN) for feature extraction and classification [18], as well as Gabor filters and Bayesian approaches [19], each achieving high recognition rates, although dataset-specific limitations were identified. A range of techniques, including Gabor wavelets, Principal Component Analysis (PCA), Local Binary Patterns (LBP), and k-NN, were combined for comprehensive feature extraction and classification [20], demonstrating relatively high accuracy.

Additionally, studies have explored methods like Kernel Principal Component Analysis (KPCA) and PCA [21], Eigenfaces with Euclidean distance metrics [22], Active Shape Models with SVM and Hidden Markov Models (HMM) [23], and Biorthogonal Wavelet Entropy with Fuzzy Multiclass SVM [24], each contributing innovative approaches and achieving noteworthy accuracy on specific datasets. Deep learning techniques, including CNNs, have also been widely adopted [25] [26]. Some studies have leveraged Transfer Learning with models like VGG16 [27], investigated facial expression recognition in NIR illumination [28], and improved FER accuracy by addressing challenges like lighting changes and occlusions [29], achieving competitive results across various datasets. Additionally, an assistive technology for visually impaired individuals has been developed using CNN-based FER [30].

Moreover, novel deep learning approaches have been proposed by combining architectures like Xception and ResNet50 for enhanced feature extraction and recognition

[31]. Lastly, EfficientNet and XGBoost techniques have been employed to achieve remarkable accuracy [32], and advanced CNN architectures have been used to identify critical facial features for FER [33]. These related works collectively demonstrate the diversity of approaches and datasets within the FER domain, highlighting the ongoing efforts to improve accuracy and applicability in real-world scenarios.

A deep transfer learning-based method for automatically classifying face expressions is proposed by researchers [34]. Convolutional neural networks (CNN) and the transfer learning model VGG19 are used in this method. Modern deep learning approaches including data augmentation, fine-tuning, and optimal learning rate finding are used to train the model. On the Japanese Female Facial Expression (JAFFE) and Extended Cohn-Kanade (CK+) datasets, the suggested model obtains accuracy values of 93.7% and 94.8%, respectively. The proposed system can be used to support smart healthcare systems, surveillance systems, and online education systems in their day-to-day operations. It has been tested on an extensive database.

Researchers in [35] use StyleGAN2 to combine transfer learning and data augmentation to present a unique method for recognizing emotions in facial expressions. Traditionally, a significant number of labeled images—which are scarce—are needed to train CNN-based models from scratch. To get around this, StyleGAN2 is used to create false expression images that correspond to seven different emotions. The emotion recognition model is then trained using transfer learning using a VGG16 network. Using augmented expression photos, transfer learning yields an accuracy of 82.04%, compared to 75.10% in the CFEE dataset in experimental results.

To improve model generalization, researchers in [36] suggest a CNN-based method for face emotion recognition that also makes use of data augmentation and transfer learning approaches. The suggested strategy outperforms current state-of-the-art models when tested on several benchmark datasets, such as FER-2013, CK+, and JAFFE. The results underscore the CNN-based approach's potential for practical uses in facial emotion identification by demonstrating how well it recognizes emotions on the face. Previous research in the field of facial expression recognition (FER) has demonstrated a large number of methodologies, ranging from traditional machine learning algorithms such as support vector machines (SVM) and k-nearest neighbors (k-NN) to deep learning techniques such as convolutional neural networks (CNNs).). These studies have demonstrated high accuracy rates, reflecting the effectiveness of the proposed methodologies across different datasets. However, weaknesses such as dataset dependency, data-hungry training, and complexity have been observed. In addition, some methods that rely on hand-crafted features are unable to capture the full complexity of facial expressions, while others suffer from overfitting without careful regularization. In contrast, our approach combines the strengths of different methodologies, including data augmentation, transfer learning, and CNN-based models, to improve performance and generalization. By emphasizing practical applicability and introducing innovative methods such as StyleGAN2 to augment data and transfer learning

5. Methodology

This research methodology was designed to develop a highly accurate image classification system by integrating cutting-edge techniques across data preprocessing, Convolutional Neural Network (CNN) model design, and the incorporation of advanced transfer learning models. It commenced with meticulous dataset selection, where we carefully chose a dataset conducive to our image classification task.

We conducted a thorough examination of this dataset, elucidating its defining characteristics, such as size and specific categories, ensuring a comprehensive understanding of the data. Subsequently, we emphasized the critical role of dataset preprocessing techniques, including data augmentation, normalization, and resizing, to prepare the data effectively for classification, enhancing its quality and diversity.

Our CNN architecture design was guided by principles encompassing convolutional layers, pooling layers, and fully connected layers, meticulously crafted to capture intricate features within images and facilitate accurate classification. A significant aspect of our methodology involved the integration of transfer learning models, wherein we provided detailed accounts of the implementation of pre-trained models such as VGG16, Inception V3, Mobile-Net, and DenseNet121. Each pre-trained model was adapted and fine-tuned to suit our specific classification task, harnessing their unique contributions. Our experimental setup was meticulously described, including the tools and techniques employed, such as Python programming language, TensorFlow, and PyTorch frameworks, as well as evaluation metrics utilized to assess model performance. In summary, our research methodology aimed to offer readers a comprehensive understanding of our experimental approach, ensuring transparency and reproducibility in our pursuit of exceptional accuracy in object recognition through image classification .

A. Overall System Block Diagram

Figure (3) presents an illustrative overview of our thesis project, outlining the high-level framework for a machine learning initiative dedicated to image classification, where a variety of deep learning models are harnessed. Let's break down each critical step in this comprehensive process.

Firstly, the crucial journey of Dataset Acquisition is embarked upon, where the dataset pivotal for training and assessing our machine learning models is procured. This dataset encompasses a diverse collection of images categorized to facilitate classification tasks, encompassing various

classes, be it objects, animals, or other pertinent entities.

The Dataset Preprocessing phase is followed, serving as a preparatory stage before model training. Here, essential tasks like standardizing image dimensions, normalizing pixel values, and possibly diversifying the dataset through augmentation techniques are engaged. This meticulous preparation ensures the dataset's readiness for subsequent model training.

Next, Dataset Partitioning is executed, with the dataset being effectively divided into distinct subsets. This typically includes the creation of a training set for model training, a validation set for hyperparameter tuning, and a testing set for comprehensive model evaluation. This partitioning strategy optimizes the model development process.

The pivotal decision of Model Selection comes into play, wherein careful choices are made from a roster of deep learning models, including CNN, VGG16, Inception V3, MobileNet, and DenseNet121, renowned for their effectiveness in image classification. Each model brings its unique architectural nuances and performance characteristics to the table.

With our models selected, Model Training is advanced through rigorous training, involving the sequential processing of images, loss calculation, and iterative weight adjustments through backpropagation. This iterative training process continues until the desired level of performance is achieved by the models.

Post-training, the models are subjected to thorough Model Evaluation, with the designated testing dataset being employed. Here, the models are put to the test, and predictions on unseen images are made, with their performance being meticulously assessed. This evaluation provides valuable insights into how effectively each model generalizes to new, unseen data.

Finally, a Performance Report is generated, encapsulating essential metrics like accuracy, precision, recall, and F1-score, tailored to the specific classification problem. These metrics collectively offer a comprehensive evaluation of each model's ability to accurately classify various classes and handle false positives and false negatives. In summary, Figure (3) lays the groundwork for a systematic and robust image classification system, integrating data preprocessing, deep learning model selection and training, and meticulous performance evaluation. This holistic framework ensures the development of highly effective models tailored to the unique demands of our image classification task.

B. *Python Colab Tool*

Python Colab, short for Google Colaboratory, is an innovative cloud-based platform that enables users to write, run, and share Python code seamlessly. Developed by Google, Colab provides a collaborative environment where multiple users can work on the same notebook simultaneously, fostering teamwork and facilitating knowledge exchange. One of its most compelling features is its integration with Google Drive, allowing users to store and access their notebooks effortlessly. Moreover, Colab offers free access to powerful GPUs and TPUs, making it an ideal choice for machine learning and data analysis tasks that demand significant computational resources. With its user-friendly interface and extensive support for popular Python libraries, Colab has become a preferred choice for educators, researchers, and professionals alike, revolutionizing the way Python code is written and executed in the cloud.

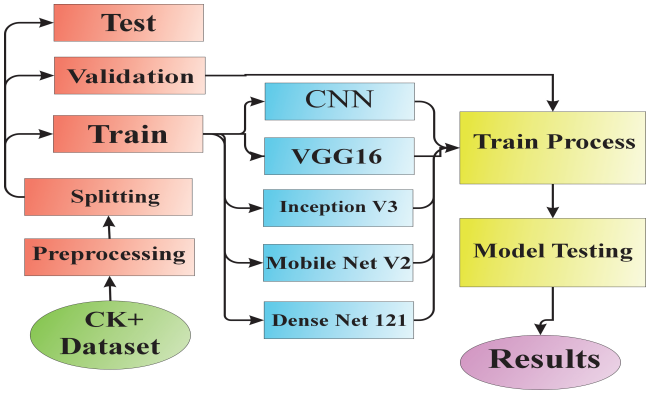

Figure 3. **System Overall Block Diagram**

C. *Cohn-Kanade Plus (CK+) Dataset Description*

This dataset comprises a comprehensive collection of facial images categorized into seven distinct emotion classes: "disgust," "contempt," "surprise," "sadness," "happy," "anger," and "fear." The dataset includes varying numbers of images for each emotion class, with 177 images depicting expressions of disgust, 54 images capturing expressions of contempt, 249 images showcasing expressions of surprise, 84 images portraying expressions of sadness, 207 images reflecting expressions of happiness, 135 images representing expressions of anger, and 75 images conveying expressions of fear. This dataset has been thoughtfully curated to provide a rich and diverse set of facial expressions, making it a valuable resource for researchers and practitioners engaged in emotion recognition, computer vision, and affective computing. Researchers can utilize this dataset to train, validate, and test machine learning models, thereby advancing the development and evaluation of algorithms for various applications, including emotion classification, facial expression analysis, and human-computer interaction. A visual representation of the dataset samples is presented in Figure (4) .

D. *FER 2013 dataset*

The FER 2013 dataset, also known as the Facial Expression Recognition 2013 dataset, is a widely used dataset in the field of computer vision and machine learning. It consists of 35,887 grayscale images of size 48x48 pixels, each labeled with one of seven facial expressions: anger, disgust, fear, happiness, sadness, surprise, and neutral. This

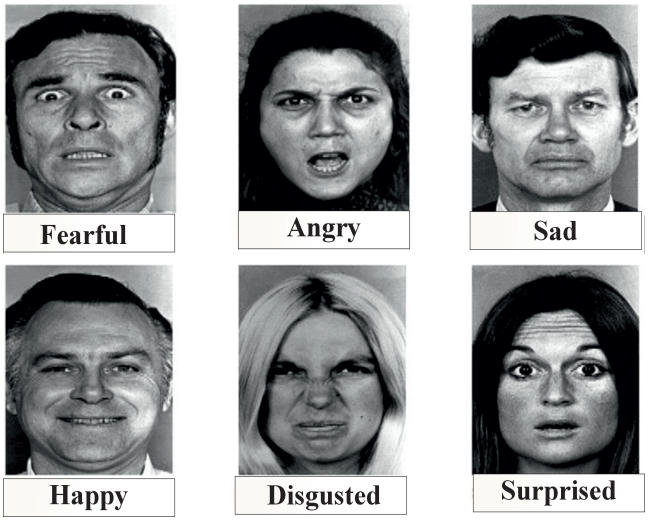

Figure 4. **. Dataset Sample** [38]

TABLE I. Dataset analysis

Emotion	Train		Test
	CK+ Dataset	FER 2013 Dataset	
Angry	108	3995	27
Disgust	141	436	36
Fear	43	4097	15
Happy	165	7215	42
Neutral	43	4965	11
Sad	67	4830	17

dataset is commonly used for training and evaluating models for facial expression recognition tasks [37]

Table (1) summarize dataset classes and numbers.

E. Dataset Preprocessing

The incorporation of an Image Data Generator assumes a pivotal role in bolstering the resilience and diversity of the training data used for machine learning models. Through the application of data augmentation techniques, this Image Data Generator offers a potent solution to address the common challenge of having a limited pool of training samples, effectively enriching the dataset's content. In the context of the provided code snippet, the 'train_datagen' object serves the purpose of preparing the training images for a range of augmentation operations. These encompass rotational adjustments, both horizontally and vertically, as well as shear transformations and variations in zoom levels.

Additionally, the option for horizontal flipping is activated, further amplifying the dataset's diversity. This augmentation procedure proves invaluable in mitigating the risk of overfitting, as it exposes the model to slightly modified renditions of the original images. Consequently, the model is encouraged to acquire a broader grasp of features and patterns, fostering enhanced generalization capabilities. By introducing controlled variations, the Image Data Generator

TABLE II. DATA AUGMENTATION FACTORS

Parameter	**Value**
Rotation Range	15
Width Shift Range	0.15
Height Shift Range	0.15
Shear Range	0.15
Zoom Range	0.15
Horizontal Flip	True

significantly contributes to the model's capacity to generalize effectively when presented with new, unseen data. This, in turn, culminates in an uplifted model performance and heightened reliability. A comprehensive breakdown of the data augmentation factors is detailed in Table (2).

Within the provided code snippet, a pivotal data preprocessing technique referred to as data normalization takes center stage. This technique stands as a fundamental stride in the preparation of image data for the training of machine learning models. The specific focus here revolves around the variables denoted as 'X_train' and 'X_test," which individually represent the training and testing datasets consisting of image samples.

Digital images fundamentally exist as matrices comprising pixel values, with each pixel denoting the intensity levels across color channels, including red, green, and blue. Typically, these pixel values span the range from 0 to 255, encapsulating the spectrum of potential intensity variations. Nevertheless, to establish a uniform foundation for training machine learning models, standardizing the data becomes imperative.

Normalization assumes a critical role in data preprocessing for several compelling reasons. Firstly, it facilitates the convergence of training algorithms, engendering stability within the optimization processes. Secondly, it plays a vital role in enhancing overall model performance by negating the influence of varying feature scales. In essence, this normalization procedure acts as a transformative step that readies image data to be seamlessly integrated into machine learning models. As a result, the models operate with heightened efficiency and accuracy throughout the training phase, a testament to the significance of this preprocessing technique.

F. CNN Model Design

The designed CNN snippet outlines the construction of a Convolutional Neural Network (CNN) model for image classification tasks. Let's denote:

· (N) as the number of classes

· (H) as the height of the input images

· (W) as the width of the input images

· (C) as the number of channels (e.g., 3 for RGB images)

The mathematical model for the CNN can be described as follows:

1. Input Layer: The input layer takes an input image with dimensions (H times W times C).

2. Convolutional Blocks: There are three convolutional blocks, each containing:

- Convolutional Layer: This layer applies a set of filters to the input feature maps.
- Normalization Layer: This could be Batch Normalization, which normalizes the activations of the previous layer.
- Pooling Layer: This layer performs down sampling to reduce spatial dimensions.
- Dropout Layer: This layer randomly drops a fraction of the input units to prevent overfitting.

3. Flatten Layer: After the convolutional blocks, the Flatten layer reshapes the output from the convolutional layers into a 1D vector.

4. Fully Connected Layers: These layers are typically Dense layers that perform classification. They take the flattened output from the previous layer and transform it using weights to produce a class prediction.

- Activation Function: Each fully connected layer typically uses a nonlinear activation function, such as ReLU (Rectified Linear Unit).
- Normalization: Batch Normalization may be applied again for normalization.

5. Output Layer: The output layer is a Dense layer with (N) units, representing the number of classes in the classification task. It uses a softmax activation function to output probabilities for each class.

Each layer applies its specific operation to the input it receives and passes the result to the next layer until the final output layer produces class probabilities. Figure (5). Shows the built CNN model.

G. Use Transfer Learning Models

Utilizing deep learning models like VGG16, Inception V3, Mobile-Net, and DenseNet121 in face emotion recognition tasks has become increasingly popular due to their exceptional capabilities in handling complex visual data. These models are adept at automatically extracting hierarchical features from images, making them well-suited for the nuanced task of recognizing and categorizing facial expressions.

1) The VGG16 model, known for its deep architecture, excels in capturing intricate patterns within facial expressions. Its extensive depth allows it to learn hierarchical features, making it effective at recognizing subtle changes in emotions. VGG16 can be fine-tuned on a dataset of facial expressions, adapting its pre-trained weights to specialize in recognizing various emotional states accurately [38].
2) Inception V3, on the other hand, is renowned for its efficiency and accuracy. It can effectively distinguish fine-grained details in facial expressions, making it an excellent choice for recognizing subtle emotional cues. Its architecture includes inception modules that enhance feature extraction, enabling it to capture complex facial expressions with precision [39].
3) Mobile-Net is a lightweight model designed for efficient processing on resource-constrained devices. Its speed and accuracy make it an attractive option for real-time face emotion recognition applications. Mobile-Net V2, in particular, can achieve impressive accuracy while requiring minimal computational resources, making it suitable for applications on mobile devices or embedded systems[40].
4) DenseNet121, a densely connected network, is another powerful choice for face emotion recognition. Its densely connected layers enable it to learn rich feature representations, making it effective at capturing and recognizing facial expressions accurately. While it may have slightly fewer parameters than some other models, DenseNet121 maintains reliable performance in this context [40].

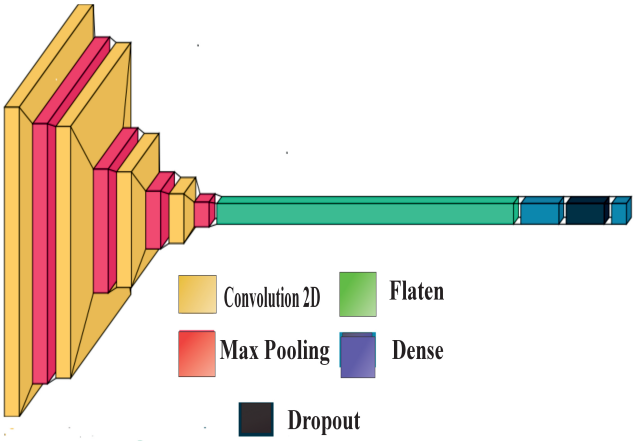

Figure 5. **Built CNN Model**

These models driven by their ability to effectively capture and recognize facial expressions, considering factors such as model depth, efficiency, accuracy, and suitability for deployment in real-world applications.

In summary, each of these deep learning models brings unique strengths to the table when it comes to face emotion recognition. Researchers and developers can choose the model that best aligns with their specific application re-

TABLE III. TRAINING CONFIGURATION SUMMARY FOR VARIOUS MODELS

Model	Optimizer	Learning Rate	Batch Size	No. of Epochs
CNN	Adam	0.001	32	30
VGG16	Adam	0.001	32	30
Inception V3	Adam	0.001	32	10
MobileNet V2	Adam	0.001	32	10
DenseNet 121	Adam	0.001	32	10

quirements, balancing factors like accuracy, computational efficiency, and model size to create effective and robust facial expression recognition systems.

H. Training Configuration Summary for Various Models

Table (3) provides a comprehensive overview of the training configurations employed across various models, encompassing Convolutional Neural Network (CNN), VGG16, Inception V3, MobileNet V2, and DenseNet121. The table succinctly presents crucial details such as the optimizer type, the chosen metric for training evaluation, the default learning rate, batch size, and the number of training epochs dedicated to each model's development. This tabulated data serves as a valuable resource for gaining insights into the specific training conditions and parameters that played a pivotal role in shaping these models' performance and capabilities.

I. Evaluation Metrics

Machine learning evaluation metrics are essential tools used to assess the performance of machine learning models. These metrics provide quantitative measures that help researchers and practitioners understand how well a model is performing on a given task. Proper evaluation is crucial in selecting the best model for a specific problem, tuning hyper-parameters, and comparing different algorithms. Table (4) describe Metrics used in this paper.

So, True Positives (TP), True Negatives (TN), False Positives (FP), False Negatives (FN).

6. Results

A. Models Accuracies

Table (5) present accuracy scores for the various algorithms used in face expression recognition offer valuable insights into their performance during the training, validation, and testing phases. Firstly, the Convolutional Neural Network (CNN) achieved commendable results with a training accuracy of 97.83%, a validation accuracy of 97.97%, and an impressive test accuracy of 98%. This indicates that the CNN model exhibits consistent and accurate recognition of facial expressions across all phases.

Moving on to the VGG16 model, while it achieved a slightly lower training accuracy of 93.24%, it demonstrated robust performance in the validation phase with an accuracy of 96.59%. In the final testing phase, VGG16 achieved a commendable accuracy of 97%. This highlights the effectiveness of VGG16, particularly in generalizing its recognition capabilities from the validation set to the test set.

TABLE IV. Evaluation Metrics

Variable	Definition
Accuracy	the percentage of accurately anticipated data from tests is easily determined by dividing all accurate forecasts by all predictions.
Precision	the proportion of outstanding instances among all anticipated ones from a specific class
Recall	the ratio of the total number of occurrences to the proportion of instances that were supposed to be members of a class
F1-Score	The phrase is used to describe a test's accuracy. The maximum F1-score is 1, which denotes outstanding recall and precision, while the lowest F1-score is 0.
Loss	This loss function is typically used in multi-class classification tasks, where is a binary indicator (0 or 1) if class label is the correct classification for observation and is the predicted probability that observation belongs to class m is the number of classes.

Inception V3 exhibited remarkable results, boasting a perfect training accuracy of 100%. During validation, it maintained high accuracy at 98.98%, and in the test phase, it achieved an outstanding accuracy of 99%. These results underline the model's exceptional ability to capture and distinguish intricate patterns within facial expressions, making it a robust choice for emotion recognition.

MobileNet V2, known for its efficiency and speed, showcased outstanding performance across the board. It achieved a flawless 100% accuracy during both training and validation, and its excellence extended to the testing phase with a perfect 100% accuracy. This demonstrates MobileNet V2's capability to recognize facial expressions with exceptional precision while being computationally efficient.

DenseNet121, although slightly lower in training accuracy at 98.72%, demonstrated competitive results during validation with an accuracy of 96.45%. However, in the final testing phase, it achieved an accuracy of 96%. Despite this slightly lower accuracy, DenseNet121 still proves to be

TABLE V. MODELS ACCURACY

Algorithm	**Train Accuracy**	**Validation Accuracy**	**Test Accuracy**
CNN	0.9783	0.9797	0.98
VGG16	0.9324	0.9659	0.97
Inception V3	1.00	0.9898	0.99
MobileNet V2	1.00	1.00	1.00
DenseNet 121	0.9872	0.9645	0.96

a reliable choice for facial expression recognition.

In summary, these accuracy scores provide valuable guidance for selecting the most suitable algorithm for specific applications, taking into account the balance between accuracy and computational efficiency. MobileNet V2 and Inception V3, with their perfect or near-perfect accuracy, stand out as strong contenders for precise and efficient facial expression recognition. Nonetheless, all models displayed robust capabilities, offering flexibility for various use cases and research contexts. Further optimization and fine-tuning may enhance their performance, making them even more valuable in real-world applications.

B. Models Loses

Table (6) provid loss values for the different algorithms used in face expression recognition offer valuable insights into the training and generalization capabilities of each model. Loss is a critical metric in machine learning that quantifies the dissimilarity between predicted and actual values, with lower values indicating better model performance.

The Convolutional Neural Network (CNN) model displayed impressive results with a minimal training loss of 0.0536 and a validation loss of 0.0764. These low loss values indicate that the CNN model effectively minimized the error between its predictions and the actual facial expression labels during both training and validation. This suggests that the CNN model learned to capture essential features and patterns in the data, resulting in accurate predictions.

VGG16, while still achieving competitive results, exhibited slightly higher training and validation losses compared to CNN. The training loss was 0.2158, and the validation loss was 0.1963. These values, while higher than those of the CNN model, are still relatively low, indicating that VGG16 successfully minimized prediction errors. However, the slightly higher losses suggest that VGG16 may not generalize as well as the CNN model.

Inception V3 performed exceptionally well in terms of loss, with a training loss of 0.0553 and a validation loss of 0.0833. These low loss values demonstrate the model's ability to accurately predict facial expressions and

TABLE VI. Loss Analysis

Algorithm	Train Loss	Validation Loss
CNN	0.0536	0.0764
VGG16	0.2158	0.1963
Inception V3	0.0553	0.0833
MobileNet V2	0.0081	0.0148
DenseNet121	0.1045	0.1433

generalize effectively. The minimal discrepancy between the training and validation losses indicates that Inception V3 is robust and capable of maintaining its performance on unseen data.

MobileNet V2 achieved the lowest loss values among all models, with a training loss of 0.0081 and a validation loss of 0.0148. These extremely low loss values highlight the model's remarkable ability to minimize prediction errors during training and validation. MobileNet V2's efficiency in learning and generalizing from the data is evident, making it an excellent choice for precise and computationally efficient facial expression recognition.

DenseNet121, while exhibiting slightly higher losses compared to MobileNet V2, still demonstrated competitive performance with a training loss of 0.1045 and a validation loss of 0.1433. These values indicate that DenseNet121 effectively learned and generalized from the training data, albeit with slightly higher errors compared to some other models.

In summary, the loss values provide a complementary perspective to the accuracy analysis, offering a detailed view of how well each model minimized prediction errors during training and validation. MobileNet V2 and Inception V3, with their remarkably low losses, stand out as strong candidates for precise and efficient facial expression recognition. However, all models showcased robust learning capabilities, catering to various application scenarios and research contexts. Further optimization and fine-tuning can potentially enhance their performance, making them even more valuable in real-world applications.

C. Results Analysis

After compute confusion matrixes, the precision, recall, and F1-score metrics offer a comprehensive evaluation of the classification performance of each algorithm in facial expression recognition. These metrics provide insights into the algorithms' ability to correctly identify specific emotional expressions while minimizing false positives and false negatives.

In terms of precision, which measures the ratio of correctly predicted positive instances to the total predicted positive instances, all algorithms performed exceptionally well. Inception V3 and MobileNet V2 achieved the highest precision scores of 0.99, indicating their proficiency in accurately identifying emotional expressions without gener-

TABLE VII. METRICS ANALYSIS

Algorithm	Precision	Recall	F1-Score
CNN	0.98	0.98	0.98
VGG16	0.97	0.97	0.97
Inception V3	0.99	0.99	0.99
MobileNet V2	0.99	0.99	0.99
DenseNet121	0.97	0.96	0.96

ating many false positives. CNN, VGG16, and DenseNet121 also demonstrated strong precision scores of 0.98 and 0.97, showcasing their precision in recognizing emotional states.

The recall metric, which quantifies the ability of the models to identify actual positive instances, echoed the strong performance of all algorithms. Inception V3, MobileNet V2, and CNN achieved recall scores of 0.99, indicating their high capability to capture true positive instances effectively. VGG16 and DenseNet121 exhibited slightly lower but still commendable recall scores of 0.97 and 0.96, respectively, demonstrating their proficiency in recognizing genuine emotional expressions.

The F1-score, which balances precision and recall, further underscores the robustness of these algorithms. Inception V3, MobileNet V2, and CNN achieved outstanding F1-scores of 0.99, signifying a harmonious balance between precise identification and recall of emotional expressions. VGG16 and DenseNet121 also delivered competitive F1-scores of 0.97 and 0.96, demonstrating their effectiveness in achieving a balance between precision and recall.

Overall, the precision, recall, and F1-score metrics collectively indicate that all the evaluated algorithms excel in facial expression recognition. Inception V3 and MobileNet V2 stand out with near-perfect scores across these metrics, making them prime candidates for applications requiring highly accurate and reliable emotional expression analysis. Nevertheless, CNN, VGG16, and DenseNet121 also demonstrate strong performance, offering a range of choices for developers and researchers depending on specific use cases and computational efficiency considerations. Table (7) explain these metrics.

These metrics affirm the effectiveness of deep learning algorithms in capturing and recognizing emotional expressions with precision and accuracy, holding great promise for a wide array of real-world applications.

D. Related works Comparison

Table (8) illustrates the varying degrees of accuracy achieved by different algorithms and methods when applied to the CK and CK+ datasets. The results highlight the effectiveness of deep learning models like CNN, VGG16, Inception V3, MobileNet V2, and DenseNet121 in achieving high accuracy in facial emotion recognition tasks. These accuracies depend on factors such as the dataset, preprocessing techniques, and model architectures, emphasizing

TABLE VIII. RELATED WORKS COMPARISON

Cite	Dataset	Algorithm	Accuracy
[17]	CK+	Hybrid	98.30%
		Transfer Learning	90.10%
	CK+	KPCA, PCA	KPCA: 76.5%, PCA: 72.3%
[23]	CK+	SVM, HMM	SVM: 70.6%, HMM: 65.2%
[43]	CK	SVM	84.68%,
[25]	CK+	CNN	C80.303%
[26]	CK+	CNN	96.76%,
Present paper	CK+	CNN	98%
Present paper	CK+	Inception V3	99%
		MobileNet V2	100%
		DenseNet121	96%

the importance of selecting appropriate methods for specific applications. Although it has a lot of promise, the suggested Convolutional Neural Network (CNN) method for picture categorization has a few drawbacks. These include issues with the scarcity of training data, the risk of overfitting, computational complexity, and sensitivity to hyperparameters.

Future research could look into integrating attention mechanisms for better interpretability, examining multimodal fusion techniques for better comprehension of complex scenes, investigating semi-supervised learning approaches to make use of unlabeled data, improving adversarial robustness, and investigating domain adaptation techniques for better real-world generalization in order to address these limitations. Through tackling these obstacles and investigating new avenues for investigation, the suggested CNN methodology can be improved and expanded upon to attain exceptional outcomes in picture classification assignments.

7. Conclusions

In this study, we investigated the performance of several deep learning models, including CNN, VGG16, Inception V3, MobileNet V2, and DenseNet121, for facial expression recognition. Leveraging two widely used datasets, the CK+ Dataset and the FER 2013 Dataset, we aimed to assess the models' precision, recall, and F1-score in accurately classifying facial expressions across various emotion classes.

The findings reveal the remarkable performance of these deep learning models in recognizing facial expressions, with minor performance variations observed among them.

Notably, Inception V3 and MobileNet V2 emerged as top performers, showcasing their superiority in this application. These models demonstrated effectiveness across different emotion classes, highlighting their robustness and versatility in facial expression recognition tasks.

However, our analysis also identified areas for improvement, particularly in terms of recall for DenseNet121. This suggests the need for further optimization and fine-tuning of the model to enhance its performance, particularly in capturing subtle nuances in facial expressions across different emotion classes.

Moving forward, future research should focus on several key areas to maximize the performance of deep learning models for facial expression recognition. Hyperparameter tuning, advanced data augmentation techniques, ensemble methods, and transfer learning approaches offer promising avenues for enhancing model accuracy and generalization.

Additionally, efforts to improve model interpretability, explore hardware optimization techniques, expand the diversity of the dataset, and test the models in cross-domain applications will contribute to advancing the field of deep learning and its practical applications in facial expression recognition.

In conclusion, this study provides valuable insights into the effectiveness of different deep learning models for facial expression recognition. By systematically evaluating these models across diverse datasets and emotion classes, we contribute to the growing body of knowledge in this field and lay the foundation for future research aimed at further improving the accuracy and applicability of facial expression recognition systems in real-world scenarios .

8. Acknowledgment

Acknowledging support from institutions such as Al Mustansiriya University is a common practice in academic works. Including a note of thanks to the university in the acknowledgments section of a paper, book, or any other scholarly work demonstrates appreciation for their support, whether it be through funding, resources, or other forms of assistance.

. References

[1] M. A. Islam, "Comparative analysis of pre-trained models and interpolation for facial expression recognition," *Metrop. Univ. Appl. Sci. Master Eng. Inf. Technol.*, 2023.

[2] S. Minaee, M. Minaei, and A. Abdolrashidi, "Deep-emotion: Facial expression recognition using attentional convolutional network," *Sensors*, vol. 21, no. 9, p. 1–16, 2021.

[3] G. Pons and D. Masip, "Supervised committee of convolutional neural networks in automated facial expression analysis," *IEEE Trans. Affect. Comput.*, vol. 9, no. 3, p. 343–350, 2018.

[4] P. Ekman, "Should we call it expression or communication?" *Innov. Eur. J. Soc. Sci. Res.*, vol. 10, no. 4, p. 333–344, 1997.

[5] G. Krishna and S. K. N. V, "Micro-expression extraction for lie detection using eulerian video (motion and color) magnification," *BLEKINGE Inst. Technol.*, 2014.

[6] P. Ekman, "Darwin's contributions to our understanding of emotional expressions," *Philos. Trans. R. Soc. B Biol. Sci.*, vol. 364, no. 1535, p. 3449–3451, 2009.

[7] F. Nonis, N. Dagnes, F. Marcolin, and E. Vezzetti, "3d approaches and challenges in facial expression," *Appl. Sci.*, p. 1–33, 2019.

[8] J. Kaur, J. Saxena, J. Shah, Fahad, and S. P. Yadav, "Facial emotion recognition," no. 1, p. 528–533, 2022.

[9] M. F. Alsharekh, "Facial emotion recognition in verbal communication based on deep learning," *Nature*, vol. 29, no. 7553, p. 1–73, 2016.

[10] J. X. Y. Lek and J. Teo, "Academic emotion classification using fer: A systematic review," *Hum. Behav. Emerg. Technol.*, vol. 2023, 2023.

[11] X. Wang and X. Wang, "Unsupervised domain adaptation with coupled generative adversarial autoencoders," *Appl. Sci.*, vol. 8, no. 12, 2018.

[12] M. Iman, H. R. Arabnia, and K. Rasheed, "A review of deep transfer learning and recent advancements," *Technologies*, vol. 11, no. 2, 2023.

[13] Z. Bala *et al.*, "Transfer learning approach for malware images classification on android devices using deep convolutional neural network," *Procedia Comput. Sci.*, vol. 212, no. C, p. 429–440, 2022.

[14] M. Solís and L.-A. Calvo-Valverde, "A proposal of transfer learning for monthly macroeconomic time series forecast," *Eng. Process.*, p. 58, 2023.

[15] A. Hosna, E. Merry, J. Gyalmo, Z. Alom, Z. Aung, and M. A. Azim, "Transfer learning: a friendly introduction," *J. Big Data*, vol. 9, no. 1, 2022.

[16] Y. Nan, J. Ju, Q. Hua, H. Zhang, and B. Wang, "A-mobilenet: An approach of facial expression recognition," *Alexandria Eng. J.*, vol. 61, no. 6, p. 4435–4444, 2022.

[17] S. A. P. Raja Sekaran, C. P. Lee, and K. M. Lim, "Facial emotion recognition using transfer learning of alexnet," p. 170–174, 2021.

[18] Y. Piparsaniyan, V. Sharma, and K. Mahapatra, "Robust facial expression recognition using gabor feature and bayesian discriminating classifier," 2014.

[19] E. Owusu, Y. Zhan, and Q. R. Mao, "A neural-adaboost based facial expression recognition system," *Expert Syst. Appl.*, vol. 41, no. 7, p. 3383–3390, 2014.

[20] D. K. Hu, A. S. Ye, L. Li, and L. Zhang, "Recognition of facial expression via kernel pca network," *Appl. Mech. Mater.*, vol. 631–632, p. 498–501, 2014.

[21] A. De, A. Saha, and M. C. Pal, "A human facial expression recognition model based on eigen face approach," *Procedia Comput. Sci.*, vol. 45, p. 282–289, 2015.

[22] M. Suk and B. Prabhakaran, "Real-time facial expression recogni-

tion on smartphones," in *2015 IEEE Winter Conf. Appl. Comput. Vision, WACV 2015*, 2015, p. 1054–1059.

[23] Y. D. Zhang *et al.*, "Facial emotion recognition based on biorthogonal wavelet entropy, fuzzy support vector machine, and stratified cross validation," *IEEE Access*, vol. 4, no. c, p. 8375–8385, 2016.

[24] K. Slimani, M. Kas, Y. El Merabet, R. Messoussi, and Y. Ruichek, "Facial emotion recognition," *Int. J. Adv. Res. Comput. Eng. Technol.*, vol. 7, no. 11, p. 88–94, 2018.

[25] A. K. Katsaggelos, S. Bahaadini, and R. Molina, "Audiovisual fusion: Challenges and new approaches," *Proc. IEEE*, vol. 103, no. 9, p. 1635–1653, 2015.

[26] S.-H. Wang, W. Yang, Z. Dong, P. Phillips, and Y. Zhang, "Facial emotion recognition via discrete wavelet transform, principal component analysis, and cat swarm optimization," *Lect. Notes Comput. Sci.*, vol. 10559, p. 203–214, Sep. 2017.

[27] H. A. Akhand, H. Sharif, S. Uyaver, J. Shao, Q. Cheng, and T. Shimamura, "Facial emotion recognition using transfer learning in the deep cnn," *Electronics*, vol. 2334, no. March, p. 549–559, 2021.

[28] C. C. Atabansi, T. Chen, R. Cao, and X. Xu, "Transfer learning technique with vgg-16 for near-infrared facial expression recognition," in *J. Phys. Conf. Ser.*, vol. 1873, no. 1, 2021.

[29] U. Dudekula and N. Purnachand, "Linear fusion approach to convolutional neural networks for facial emotion recognition," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 25, no. 3, p. 1489–1500, 2022.

[30] D. Shehada, A. Turky, W. Khan, B. Khan, and A. Hussain, "A lightweight facial emotion recognition system using partial transfer learning for visually impaired people," *IEEE Access*, vol. 11, p. 36961–36969, 2023.

[31] M. P. Sunil and S. A. Hariprasad, "Facial emotion recognition using a modified deep convolutional neural network based on the concatenation of xception and resnet50 v2," *SSRG Int. J. Electron. Commun. Eng.*, vol. 10, no. 6, p. 94–105, 2023.

[32] S. B. Punuri *et al.*, "Efficient net-xgboost: An implementation for facial emotion recognition using transfer learning," *Mathematics*, vol. 11, no. 3, p. 1–24, 2023.

[33] Z. Y. Huang *et al.*, "A study on computer vision for facial emotion recognition," *Sci. Rep.*, vol. 13, no. 1, p. 1–13, 2023.

[34] A. Sultana, S. Dey, and M. A. Rahman, "Facial emotion recognition based on deep transfer learning approach," *Multimed. Tools Appl.*, vol. 82, May 2023.

[35] T. Kusunose, X. Kang, K. Kiuchi, R. Nishimura, M. Sasayama, and K. Matsumoto, "Facial expression emotion recognition based on transfer learning and generative model," in *2022 8th International Conference on Systems and Informatics (ICSAI)*, 2022, p. 1–6.

[36] A. Pradesh, V. P. Kowshik, C. N. Kalyani, and K. M. Teja, "Facial emotion detection using convolutional neural network," *J. Eng. Sci.*, vol. 14, no. 04, 2023.

[37] L. Zahara, P. Musa, E. Prasetyo, I. Karim, and S. Musa, "The facial emotion recognition (fer-2013) dataset for prediction system of micro-expressions face using the convolutional neural network (cnn) algorithm based raspberry pi," 2020.

[38] R. P. Holder and J. R. Tapamo, "Improved gradient local ternary patterns for facial expression recognition," *Eurasip J. Image Video Process.*, vol. 2017, no. 1, 2017.

[39] S. Wu, "Expression recognition method using improved vgg16 network model in robot interaction," *J. Robot.*, vol. 2021, p. 9326695, 2021.

[40] A. Agarwal and S. Susan, "Emotion recognition from masked faces using inception-v3," 2023.

Raed Ibrahim Khaleel Almsari A dedicated researcher, he began his academic journey with a strong foundation in electrical engineering. He obtained a bachelor's degree. He is currently a graduate student at Al-Mustansiriya University in Iraq, demonstrating early promise in this field. Based on this achievement, Raed pursued his passion for control and computer engineering and obtained a master's degree.

Abbas Hussien Miry received his B.Sc. degree in Electrical Engineering in 2005 from the Mustansiriyah University and his M.Sc. degree in control and computer engineering in 2007 from Baghdad University. He received a Ph.D. degree in 2011 in control and computer engineering from the Basrah University, Iraq. In 2007. His recent research activities are artificial intelligence, control and swarm optimizations. He can be contacted at email: abbasmiry83@uomustansiriyah.edu.iq .

Tariq M. Salman was born in Baghdad, Iraq in 1972. He obtained his B.Sc. in Electrical Engineering in 1995, M.Sc. in Communication Engineering in 2003 at University of Technology, Iraq, and Ph.D. in Telecommunication and Network Devices in 2012 at Belarussian State University of Informatics and Radio Electronics, Belarus. From 2006 to 2012, he worked as a lecturer in the Electrical Engineering Faculty, at Al-Mustansiriyah University, Iraq. Since the beginning of 2018, he has worked as an assistant professor in the same Faculty. He has been a consultant mem- ber of the Iraqi Engineering Union since 2013. He is interested in the subject of wireless and network devices, video, and image processing systems

email: tariq. salman@uomustansiriyah.edi.iq