Dicle University
Journal of Engineering
https://dergipark.org.tr/tr/pub/dumf
duje.dicle.edu.tr

Research Article

Application of Emotion Analysis in Deep Learning Techniques

Mesut UYSAL[1*], Mehmet Fatih DEMİRAL[2], Ali Hakan Işık[3]

[1] Hatay Mustafa Kemal University, Antakya Vocational School, Department of Database Network Design and Management, mesut.uysal@mku.edu.tr, Orcid No: 0009-0002-1650-8880
[2] Burdur Mehmet Akif Ersoy University, Department of Data Science and Analytics, mfdemiral@mehmetakif.edu.tr Orcid No 0000-0003-0742-0633
[3] Burdur Mehmet Akif Ersoy University, Computer Engineering Department, ahakan@mehmetakif.edu.tr, Orcid No: 0000-0003-3561-9375

ARTICLE INFO

Article history:

Received 3 August 2025
Received in revised form 13 October 2025
Accepted 25 November 2025
Available online 30 December 2025

Keywords:

Emotion recognition, Deep learning, Transfer learning, Facial expression, YOLOv8m-cls

Doi: 10.24012/dumf.1757225

* Corresponding author

ABSTRACT

Emotion recognition has become a pivotal technology in advancing human-computer interaction with applications spanning fields such as healthcare, entertainment, and customer experience. This paper evaluates the performance of five deep learning models—YOLOv8m-cls, ResNet50, EfficientNetB5, MobileNetV2, and DenseNet121—in detecting emotions from facial expressions. Leveraging the AffectNet dataset, which initially contained eight emotional categories, we focused on five emotions after excluding three due to low data availability and similarity. The emotions processed include anger, happiness, sadness, surprise, and fear. The models were fine-tuned through transfer learning, demonstrating that YOLOv8m-cls performed best, balancing accuracy, speed, and generalization, making it suitable for real-time applications. ResNet50 and EfficientNetB5 also performed well, with ResNet50 excelling in handling complex facial features and EfficientNetB5 offering computational efficiency with high accuracy. The study also highlights challenges such as intra-class variability and inter-class similarity, which continue to affect model performance. These findings underscore the importance of selecting model architectures based on specific application requirements and suggest that future research should explore integrating multimodal data to enhance emotion recognition systems.

Introduction[1]

Emotion recognition has emerged as a critical component in advancing human-computer interactions and improving how machines understand human behavior. At the core of emotion recognition is the challenge of deciphering complex emotional states that involve a mix of cognitive, physiological, and behavioral responses. Emotions such as happiness, sadness, anger, fear, surprise, and disgust are often communicated through a variety of channels, including facial expressions, vocal tones, and body language. The ability to accurately identify these emotions has far-reaching implications not only for improving user experiences with technology, but also for advancing fields such as mental health, social robotics, and emotional computing. As one of the most direct and universal indicators of emotion, facial expressions have been a primary focus of emotion recognition systems. Across cultures and contexts, facial expressions serve as a powerful medium for conveying emotional states [1]. These expressions consist of complex facial muscle movements that can be deciphered to reveal the underlying emotion. Early research on facial expressions for emotion recognition relied largely on manual annotations and rule-based systems. Such systems used hand-crafted features that tracked specific facial cues and mapped them to corresponding emotions. However, these methods were often limited in their ability to adapt to dynamic environments, such as changes in lighting, occlusions, or differences in individual facial features [2].

The practical applications of emotion recognition extend beyond simple human-computer interactions. In healthcare, emotion-detection systems are increasingly being used in mental health diagnostics to monitor emotional states in patients and to help diagnose disorders such as depression, anxiety, and post-traumatic stress disorder (PTSD). In marketing and customer service, companies use emotion recognition to better understand consumers' responses to products or advertisements. These insights can lead to more personalized marketing campaigns and increased customer satisfaction. Emotion recognition also plays an important role in entertainment, particularly in the development of emotionally responsive characters in video games and virtual reality environments [3].

[1] This work is based on the master thesis of Uysal(2024)

Over the past two decades, emotion recognition has evolved from simple rule-based systems to more complex, data-driven models. This evolution has been largely driven by advances in artificial intelligence (AI) and machine learning (ML), particularly computer vision. By leveraging large datasets and powerful computational models, modern emotion recognition systems can automatically learn and detect features that correspond to emotional states without the need for extensive manual intervention [4]. Such systems have demonstrated remarkable accuracy in identifying emotions across a wide range of conditions, making them highly adaptable and reliable. As one of the main data sources for emotion detection, face recognition continues to dominate the research landscape. Automatic face recognition systems analyze pixel-level data from images or video streams to identify patterns in facial landmarks and textures that correspond to specific emotions. These systems typically rely on sophisticated feature extraction techniques, where key facial landmarks, such as the position of the eyes, mouth, and eyebrows, are tracked and processed using deep learning architectures. However, emotion recognition is not without its challenges. Variability in facial expressions due to cultural differences, individual characteristics, and environmental factors (such as lighting and occlusion) can significantly affect the accuracy of these systems [5]. Despite all these difficulties, this study aimed to perform accurate emotion detection on 29,042 images taken from the Kaggle platform by selecting transfer learning models. The performance of each model was measured using various metrics, and the model with the highest accuracy value was selected and its suitability for this study was indicated.

Literature review

Previous studies on emotion detection

Emotion detection has long been a topic of research interest, evolving through various methodologies and technological advancements. The core aim of emotion detection is to interpret emotional states from various forms of data such as facial expressions, voice, text, and physiological signals. Early research on this topic largely relied on handcrafted feature extraction techniques, often based on psychological theories of emotion, such as Ekman's six basic emotions [1]. However, with the advancements in machine learning and artificial intelligence, especially deep learning, the field has experienced significant transformations, leading to more accurate and scalable emotion detection systems.

Initial studies in emotion detection focused heavily on the manual interpretation of facial expressions. Pioneering study like Ekman and Friesen introduced the Facial Action Coding System (FACS), which provided a systematic method for describing facial expressions by identifying specific facial muscle movements, known as Action Units (AUs) [6]. This system became the foundation for many early facial emotion recognition studies, as researchers sought to decode emotions through observable facial movements. However, these methods were limited in their scope and lacked the computational power to be applied on a large scale.

As computing power increased, so did the ambition of emotion detection research. Early machine learning models, such as Support Vector Machines (SVMs), Naive Bayes classifiers, and Decision Trees, were applied to automate the recognition of emotional states. These models, while offering some level of automation, relied heavily on feature engineering—human intervention to manually define the relevant features for emotion recognition where Tian, Kanade and Cohn proposed a facial expression recognition system that used feature-based SVMs to classify expressions based on geometric and appearance features from the face [5]. Although effective under controlled conditions, these models struggled with real-world variability, such as changes in lighting, head poses, and occlusions, limiting their practical applications.

The turn of the century saw a marked shift towards data-driven approaches, facilitated by the rise of large emotion-labeled datasets and advances in deep learning. One such dataset is the Extended Cohn-Kanade (CK+) dataset, which contains labeled facial expressions captured under various conditions [7]. This dataset became a benchmark for facial expression recognition and contributed to the development of more robust emotion recognition systems.

The introduction of Convolutional Neural Networks (CNNs) revolutionized the field by enabling automatic feature extraction directly from raw data, such as images and videos, without the need for manual intervention. CNNs consist of multiple layers of processing units that learn hierarchical representations of the input data, making them highly effective for image-based tasks like emotion detection. Notably, the work of Tolias and Chum demonstrated the power of CNNs in facial expression recognition [8]. Their proposed deep neural network achieved state-of-the-art performance on several datasets, including FER2013 and AffectNet, two large-scale facial emotion recognition datasets.

CNNs are designed to handle the spatial hierarchies in image data, which is crucial for identifying patterns that exist in local neighborhoods of pixels, such as the shape of a mouth or the orientation of the eyes in facial emotion recognition tasks. The network's ability to learn these spatial hierarchies without manual intervention has made it one of the most powerful tools in image classification. One of the earliest successful applications of CNNs in image classification was the AlexNet model, which won the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) in 2012 by a significant margin [9]. AlexNet’s success demonstrated the immense potential of deep learning for image classification and spurred further research into deeper and more efficient networks.

The measurement of the success of secondary school activities conducted by Zilyas and Yılmaz, and their learning by determining negative features, machine learning [10]. The recorded data set of a survey results including three questions consisting of 519 students, 246

male and 273 female students, consists of 13 columns and 520 rows. The survey asked 13 questions regarding gender, number of siblings, whether they receive special education, and whether their mother and father are separate/together. The data set; After dividing into 80% training, 20% test, data cleaning and pre-processing steps were passed in order to obtain high performance before training, and training was performed in Random Forest, Multiple Linear Regression, Gradient Boosting Regression, KNN, Bagged Trees, Decision Trees explosions. According to successful learning machine applications, the output acquisition rate is 98% with Random Forest cleaning and 95% with R-squared scores.

The FER2013 dataset, introduced during the ICML 2013 Challenges, is widely used in the emotion recognition community and contains over 35,000 labeled images of faces expressing seven different emotions. The AffectNet dataset is even larger, comprising more than one million facial images annotated with both categorical and continuous emotion labels. These large datasets have been instrumental in training deep learning models that are capable of generalizing across different emotional expressions, individuals, and environmental conditions. CNN-based models trained on these datasets have consistently achieved superior performance compared to traditional machine learning methods, making them the go-to solution for modern emotion detection tasks [11].

Despite the success of CNNs, challenges remained, particularly when dealing with sequential data such as video or audio, where temporal dynamics are crucial for understanding emotion. Recurrent Neural Networks (RNNs) and their variants, Long Short-Term Memory (LSTM) networks, emerged as suitable candidates for handling sequential data. Wöllmer et al. applied LSTMs to emotion recognition from speech, demonstrating that RNNs can capture temporal dependencies in emotional expressions, which are often missed by static image-based methods. Similarly, in video-based emotion detection, LSTMs have been used to model the temporal evolution of facial expressions over time, providing more nuanced and accurate predictions [12].

Transfer learning has further propelled the advancements in emotion detection. Transfer learning allows models pre-trained on large general-purpose datasets, like ImageNet, to be fine-tuned on smaller, domain-specific datasets. This technique has proven highly effective in emotion detection, as emotion-labeled datasets are typically much smaller than general-purpose datasets due to the complexity and cost of labeling emotions. By leveraging pre-trained models such as DenseNet, EfficientNet, MobileNet, researchers have been able to achieve high performance with fewer labeled samples [12]. Transfer learning not only reduces the need for extensive computational resources but also accelerates the training process and improves generalization to unseen data. In to facial expression recognition, emotion detection from speech and text has also seen significant progress. In speech-based emotion detection, studies such as Wöllmer et al. and Eyben et al. showed that acoustic features such as pitch, intensity, and speaking rate can be used to infer emotional state.[12-14] Similarly, text-based emotion recognition, which focuses on identifying emotions from written content, has gained traction with the rise of social media platforms. Sentiment analysis, a subset of emotion detection, is commonly used to gauge public sentiment on social media posts, product reviews, and other textual data. Natural Language Processing (NLP) techniques, including the use of pre-trained language models like BERT. Devlin et al. have significantly improved the accuracy of text-based emotion detection [15]. A growing body of research is also exploring multimodal emotion detection, which combines data from multiple modalities, such as facial expressions, speech, and text, to enhance the accuracy of emotion recognition. For instance, Zadeh et al. proposed a multimodal deep learning model that integrates visual, auditory, and textual cues to predict emotions in video data [16]. Their model outperformed unimodal approaches, highlighting the importance of considering multiple channels of emotional expression for more comprehensive emotion detection. Despite these advancements, emotion detection remains a challenging task due to the complexity of emotional states and the variability in how individuals express emotions. Factors such as cultural differences, personal traits, and situational contexts can significantly affect the accuracy of emotion recognition systems. Furthermore, ethical concerns regarding the use of emotion detection technologies, particularly in surveillance and privacy, have sparked debates about the potential misuse of such systems. As emotion detection continues to evolve, addressing these challenges will be crucial for the responsible development and deployment of emotion-aware technologies.

Shawi et al. [17] found that ResNet-50 architectural enhancements improved face emotion recognition performance. Introducing hybrid pooling with a learnable α weight and enhancing the Adaptive Leaky ReLU activation function with a parameter led to improved efficiency and flexibility in learning complicated emotional variables. In experiments on the FER2013 and CK+ datasets, the enhanced model attained accuracies of 96.60% and 96.32%, outperforming ResNet-50 alone, ResNet-50 with CBAM, and ResNet-50 with GRU. CK+ had a slightly superior precision-recall balance than FER2013, with an average F1-score of 0.9424 against 0.931. “Happiness” was the most reliably detected emotion, but “Neutral” performed poorly, especially on FER2013. This shows that adding Hybrid Pooling and Adaptive Leaky ReLU to ResNet-50 enhances feature representation and classification, making face emotion identification more accurate and dependable across complicated datasets.

A research by Alshammari, & Alshammari [18] showed that the YOLOv8 framework is a powerful model for emotional facial expression identification that can recognize many emotion categories. The model performed well on 2,353 pictures labeled with seven emotions—anger, contempt, disgust, fear, pleasure, sorrow, and surprise—with a mean Average Precision (mAP@0.5) of

0.837. Anger (1.00), pleasure (0.95), contempt (0.93), disgust (0.70), and surprise (0.88) were more reliably identified than fear (0.36) and melancholy (0.25). Fear and melancholy performed poorly owing to data imbalance, since these classes included less training examples (75 and 84 photos, respectively). YOLOv8 trained to recognize and categorize emotional expressions with increasing accuracy as its precision, recall, and mAP improved over training epochs. Scatter plot analysis also showed that the model detected centrally situated and bigger facial emotions more effectively, showing face placement and size sensitivity. The results show that YOLOv8 excels at emotion identification, but dataset augmentation and class balance might improve its detection of subtle emotions like melancholy and fear.

Istiqomah, et al. [19] showed that transfer learning using ResNet-50 improves facial emotion recognition over training from scratch. Image improvement and resizing helped the model achieve 99.49% accuracy, 99.49% precision, 99.71% recall, and 99.60% F1-score across seven emotion classes, proving its dependability. Transfer learning outperformed non-transfer learning in classification convergence, stability, and true positive rates. These results demonstrate that transfer learning enhances the model's capacity to handle illumination, viewing angles, and face structures, making it a very successful facial emotion identification method.

Wei Du [20] suggested enhancing ResNet50 for face emotion identification. The suggested ResNet50 model has two fully linked layer blocks. The layer stacking approach with reduced connections is utilized to address deterioration and enhance training flexibility. The enhanced model outperformed ResNet50 by 13.31% on the dataset (FER-2013). ResNets have downsides, such as relying significantly on Batch Normalization, which stabilizes training and lowers Dropout. In residual networks, maintaining information flow over skip links is crucial. Dropout may diminish this value.

To recognize and categorize facial expressions, Liu Luan Xiang et.al., [21] suggested a system using ResNet50 and a convolutional block attention module. Deep learning models (VGG19, ResNet50, and InceptionV3) extensively assessed FER tasks and addressed difficulties. The convolutional block attention module was introduced to improve retrieved features and enhance model performance by hyper parameter adjustment. In experiments, VGG19's accuracy was 71.7% before the module integration, whereas ResNet50's accuracy peaked at 72.4% after integration. However, there are some downsides. Adding CBAM to every ResNet layer (50) will dramatically increase calculations, resulting in: A longer training period is required to calculate attention units for each layer. Higher memory use (VRAM/RAM) occurs with huge datasets or high-resolution photos. A slower inference time may be impractical for real-time applications. Using CBAM in all layers may overemphasize individual characteristics, reducing model generalization and overfitting on fresh data. The use of CBAM at each layer makes it difficult to adjust settings and identify the impact of each attention unit, making analysis challenging.

Deep learning methods, especially hybrid CNN-RNN architectures, are best for face emotion identification because they capture both spatial and temporal qualities of facial emotions. Attention processes and embeddings do not considerably outperform RNN-based models. The study [22] found that dataset size and computing restrictions limit deep network training. Transfer learning and preprocessing methods like noise reduction and contrast enhancement improved recognition accuracy. Additionally, ensemble approaches and adversarial training may minimize mistakes by improving model resilience. Optimization is necessary for hyperparameters including learning rate, batch size, number of layers, dropout rate, and activation function, which greatly affect model performance. Finally, the study found that multimodal approaches that combine facial cues with other emotional signals like ECG or verbal cues can improve emotion detection, and that future research should focus on recognizing continuous, spontaneous, and subtle expressions in real life.

Deep learning-based approaches, particularly hybrid ConvNet-RNN and ConvNet-LSTM architectures, outperform traditional machine learning methods in facial expression recognition (FER) because they automatically extract features and capture spatial and temporal information from images and video sequences. Deep learning allows end-to-end learning from raw photos without face physics-based models like SVM, k-NN, and AdaBoost. Training deep networks is difficult due to restricted sample sizes and processing needs, hence the research emphasizes big, high-quality datasets. 3D datasets and multimodal techniques with temporal and physiological information may enhance identification accuracy, especially for minor emotions and position changes. Deep learning algorithms improve FER performance, but more effort is required to construct models that can handle complicated, real-world expressions and minimize computational cost for practical applications [23].

Literature review

In Section 2.2, the contents, data sources and results of previous studies on sentiment analysis are mentioned.

Table 1. Literature table.

Reference	Algorithm/Libraries	Contents	Data source	Language	Results
Vural et al. [**24**]	Dictionary lexicon	Movie reviews	White screen site	Java	-
Bilgin and Şentürk [**25**]	Doc2Vec, Dm, DBow	Social media comments	Twitter	Python	%34 DM precision %44 DBow precision

Demirci et al. [**26**]	NN, SVM, LR	Political news	Twitter	Python	%81 accuracy (ML) %81,86 accuracy (DL)
Seyfioğlu and Demirezen [**27**]	Semi-supervised learning	Aircraft Customer Reviews	Customer reviews and ratings	Python	%92,5 accuracy
Zilyas and Yılmaz [[10]]	LR, Random Forest, KNN, Btr	Measuring Student Success	Survey results	Python	%88 accuracy
Ayan, Kuyumcu, Ceylan [**28**]	Lineer Ridge Regresyonu ve NB	Islamophobic thoughts	Tweets about Islam on Twitter	Python	%95,4 F1 score
Uyaroğlu Akdeniz and Cebeci [**29**]	Turkish Bert	Municipality Service satisfaction	Twitter comments	Python	%80 accuracy
Tuzcu [**30**]	MLP, NB, LR, DVM	Book Sales	Book sales reviews	Python	%90 DVM accuracy %88 MLP accuracy
Yohanes et al. [**31**]	LSTM, BiLSTM, gru	Researching and grouping people's basic emotions	ISEAR data set	Python	%98,72 accuracy
Zhang and Ananiadou [**32**]	CNN models, Deep learning models	Gender effects on sentiment analysis	ISEAR ve CrowdFlower	-	%99 accuracy
Issa et al. [**33**]	Deep learning models	Identifying emotions from speech	RAVDESS, EMO-DB, IMOCAP	Python	%95,71 EMO-DB accuracy
Yüksel and Tan [**34**].	NLP, Zemberek, Simple Bayesian Algorithm	Feelings about places visited via Foursquare	Comments received from Foursquare	Python	%81 binary classification accuracy %84 triple classification accuracy
Arığ and Turan [**35**]	CNN deep learning model	Video sentiment analysis	Reviews from 61 movies	Python	%85 ML accuracy
Eyipınar et al. [**36**]	Text Mining Techniques	Healthy nutrition and sports	The 6 most watched sports videos	Orange 3	%60 accuracy
Tepecik and Demir [**37**]	Nb, SVM, RF KNN, DT	Sentiment analysis from voice	Common Voice	Python and Rapid Miner	Naive Bayes %70 accuracy SVM %69 accuracy
Sel [**38**]	N-gram method	Opinions on the pandemic	Twitter	Python	%70 accuracy
Uysal [**39**]	Deep learning models	Emotion Detection	Kaagle	Python	%89 accuracy

Methodology

Dataset description

For this study, we used a subset of the widely known AffectNet dataset, which initially consisted of facial expressions categorized into eight different emotional states: anger, contempt, disgust, fear, happiness, neutral, sadness, and surprise. However, due to limited data availability for certain emotions and the similarity between some emotion pairs, we continued with five basic emotions: anger, happiness, neutral, surprise, and sadness. The excluded emotions were disgust, contempt, and fear. The total number of images in the dataset is distributed as follows:

Total number of images in the dataset: 29042

Number of images used in models: 20518

- Anger: 3218 images
- Happiness: 5044 images
- Neutral 5126 images
- Surprise: 4039 images
- Sadness: 3091 images

The images in the dataset were converted from 96*96 to 224*224 so that the model can better analyze the images and gain an advantage in detecting the correct emotion. In addition, the images in BGR format were converted to RGB format before being transferred to the models. No data extraction, deletion or data augmentation was done while editing the dataset.

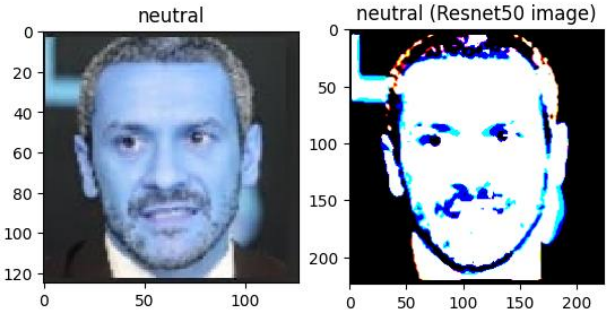
Preprocessing

AffectNet Training Data is one of the largest and most comprehensive datasets developed for facial emotion recognition studies. Approximately 450,000 images were manually labeled with eight primary emotion labels. The labeled data included Anger, Disgust, Fear, Happiness, Sadness, Surprise, Neutral, and Contempt. Approximately 29,042 images were selected from the 450,000 images without any specific criteria. Our dataset contains a total of 29,042 images, but 20,518 of these were used for training and testing the models. Because the accuracy values for predicting eight emotions were not satisfactory for the project results, we preferred to use five classes that yielded better results. Of these eight classes, the contempt and disgust classes were removed from the dataset due to limited data. Because the fear and sadness classes contained similar data, the fear class was also removed from the dataset in the next stage. The two classes with the fewest data in the dataset were removed, while the classes with the most data were used for training. The data in the dataset was already labeled, so no separate labeling was performed a second time. When creating the Affectnet dataset, the facial expressions in this dataset were manually labeled by human analysts. This labeling process is based on psychological and visual criteria. Thousands of images are collected from search engines like Google and Bing. Each image is evaluated by at least one analyst, who labels

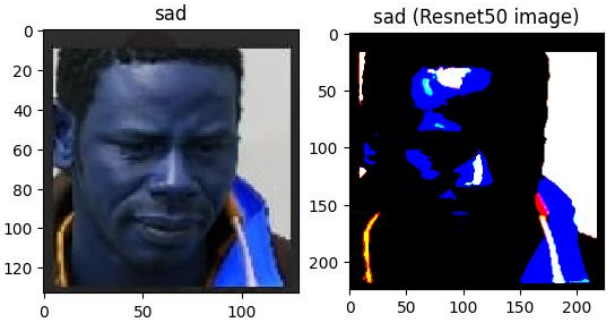
the images with continuous values such as pleasure and arousal [-1,1]. Additionally, landmarks, bounding boxes, and pose information (face angle) were used to create the Afectnet datasets.

Picture 1: Sample Images of Emotional Classes Data Set

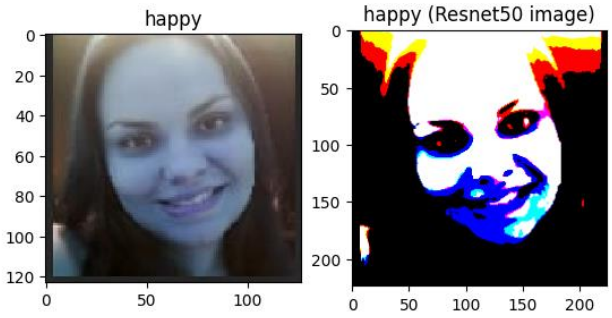
Picture 2: Images before and after preprocessing on Restnet50 (Surprise)



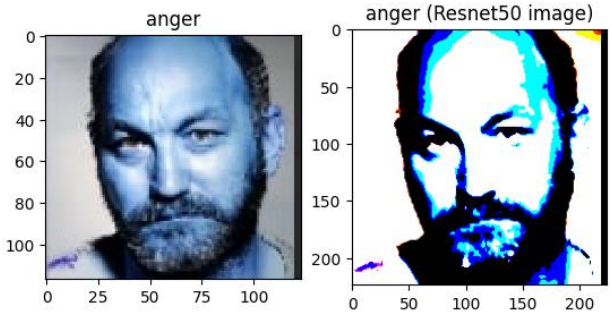
Picture 3: Images before and after preprocessing on Restnet50 (Neutral)



Picture 4: Images before and after preprocessing on Restnet50 (Sad)



Picture 5: Images before and after preprocessing on Restnet50 (Happy)



Picture 6: Images before and after preprocessing on Restnet50 (Anger)

The unbalanced representation of emotions like "hate, disgust, and fear" in the dataset created some performance issues. The model struggled to learn certain classes, leading to a decrease in overall model performance during the training process. If all datasets had been assigned equally, a completely different result could have been achieved. However, it performed quite well with happiness and neutral emotions.

Normalization

After resizing, all pixel values of the images were normalized. Image normalization is an essential step in most deep learning tasks, as it scales the pixel intensity values to a fixed range, typically between 0 and 1. In this study, normalization was performed by dividing the pixel values by 255, which is the maximum possible pixel value in an 8-bit image. This scaling ensures that the inputs to the neural network have a standardized range, helping the models learn more effectively and converge faster during training.

Splitting the dataset

Before the actual training process, the dataset was split into two parts:

- Training Set: 80% of the images were used to train the model.

- Test/Validation Set: 20% of the images were used for both testing and validation. We did not separate a dedicated validation set; instead, the test data was used concurrently for validation during the training process. This combined test/validation set helped monitor the model's performance and prevent overfitting by providing checkpoints to evaluate how well the model was generalizing to unseen data. Because the final test results may not fully reflect generalization ability, it is better to use separate test and validation sets for future studies.

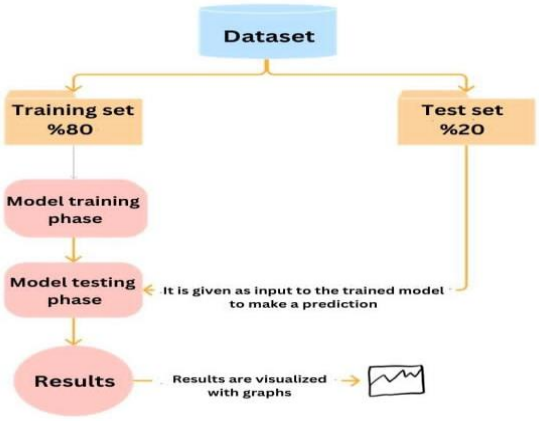


Figure 1. Data distribution percentage.

Model selection

In this section, we explore the selection of six advanced deep learning models that were employed for emotion recognition. These models—DenseNet121, EfficientNetB5, MobileNetV2, ResNet50, VGG16, and YOLOv8m-cls—have been chosen due to their proven performance in image classification tasks and their adaptability to transfer learning techniques.

DenseNet121

DenseNet121 (Densely Connected Convolutional Networks) is a powerful deep learning model that leverages dense connectivity patterns between layers. Unlike traditional convolutional neural networks, where each layer only feeds into the next one, DenseNet121 connects each layer to every other layer in a feed-forward fashion. This connectivity improves the flow of information and gradients throughout the network, addressing the vanishing gradient problem commonly encountered in deep networks [3].

DenseNet121 is well-suited for emotion recognition tasks due to its ability to capture intricate details in facial expressions. In this study, the model was fine-tuned using pre-trained weights from the ImageNet dataset and adapted to recognize the eight primary emotion classes. The input shape was set to *224x224 pixels,* and the model utilized *max pooling* and *global average pooling* layers to reduce the dimensionality of the data while retaining critical features for classification.

EfficientNetB5

EfficientNetB5 is a highly efficient and scalable deep learning model designed to balance accuracy and computational complexity. The model uses a compound scaling technique, which uniformly scales network depth, width, and resolution. This allows EfficientNetB5 to achieve superior performance with fewer parameters compared to traditional models.

In this study, EfficientNetB5 was pre-trained on ImageNet and fine-tuned for emotion recognition using the ImageNet dataset. The input images were resized to *224x224 pixels,* and a *softmax classifier* was used to output the probability distribution over the emotion classes.

MobileNetV2

MobileNetV2 is an efficient convolutional neural network optimized for mobile and embedded applications. It employs depthwise separable convolutions to reduce the computational load while maintaining high classification accuracy. MobileNetV2 is especially useful in scenarios where computational power is constrained, such as on smartphones or edge devices [40].

ResNet50

ResNet50 (Residual Networks) is a deep convolutional neural network that uses residual learning to overcome the vanishing gradient problem and enables training of very deep networks [41]. ResNet50's architecture consists of 50 layers and includes shortcut connections that skip one or more layers, which makes gradients flow more easily during training. ResNet50 was selected for this study due to its strong performance in image classification and good generalization across different datasets.

VGG16

VGG16 is a deep convolutional neural network that is widely recognized for its success in the ImageNet competition. It uses a simple and consistent architecture consisting of 16 layers consisting primarily of 3x3 convolutional filters followed by max pooling layers [42]. The simplicity of VGG16 combined with its deep architecture enables it to capture complex features in images, making it a popular choice for transfer learning tasks.

YOLOv8m-cls

YOLOv8m-cls (version 8 for You Only Look Once classification) is an advanced version of the YOLO object detection framework adapted for image classification tasks. YOLO models are renowned for their speed and accuracy in real-time object detection, making YOLOv8m-cls a suitable choice for emotion detection in dynamic environments such as video streams or live interaction systems [43]. The primary advantage of YOLOv8m-cls is its ability to process images at high speeds while maintaining high accuracy, making it ideal for real-time emotion recognition applications. In this work, the classification mode of the YOLOv8 architecture developed by Ultralytics was used for image classification.

Table 2. YOLOv8m-cls hyperparameters

Model	YoloV8m-cls
Epoch	20
Lr	0.01
Batch size	16
Optimizer	Adam
İmgsz	96

Training and evaluation

The training and evaluation process is a critical phase in the development of emotion recognition models. In this study, the chosen models (DenseNet121, EfficientNetB5, MobileNetV2, ResNet50, VGG16, and YOLOv8m-cls) were trained and evaluated using the same dataset and preprocessing steps to ensure consistency and comparability across models. The training process for each model involved fine-tuning pre-trained weights on the AffectNet dataset and adapting the models to the emotion recognition task. The key stages in training and evaluation are described below.

Training process

Each model was trained using a training set of 16,414 images, while 4,104 images were reserved for testing. Since there were not many images in the dataset, the test data was also used as validation data. In other words, no separate split was made for validation. The training process involved updating the model weights to minimize the loss function, which measures the difference between the predicted and true emotion labels. The primary loss function used in this study was categorical cross-entropy, which is suitable for multi-class classification tasks such as emotion recognition. This loss function calculates the probability that the correct label is predicted by the model and adjusts the model's weights accordingly to improve its predictions.

In training, each model was trained 30 times and processed 128 data points simultaneously. These parameters were chosen to ensure that the models had sufficient opportunity to learn the complex relationships between facial features and emotions while avoiding overfitting. Overfitting occurs when a model performs well on training data but fails to generalize to unseen data, and is a common problem in deep learning. To reduce this risk, early stopping and learning rate reduction techniques were used.

- Early Stopping: If the performance of the model on the test/validation set has not improved over 5 epochs, training is stopped early to prevent overfitting. In this study, the accuracy metric was monitored.
- Learning Rate Reduction: These metrics are monitored in the loss graph. If the losses in the model are high, the learning rate is reduced by 0.1. The parameter patience=2 specifies the number of rounds to wait before reducing the learning rate.

System architecture

The diagram shows the overall system architecture used for emotion recognition. The process starts with importing the necessary libraries, including TensorFlow, NumPy, and OpenCV. The video camera captures a frame, which is then preprocessed and passed through a deep learning model using transfer learning techniques. The model estimates the emotional state from facial expressions and displays the result on the screen. This architecture efficiently integrates real-time image processing with deep learning to provide accurate emotion detection.

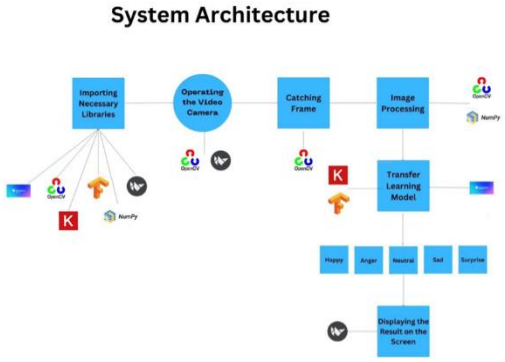


Figure 2. System architecture of the study.

Common metrics for model performance in emotion recognition

In the evaluation of emotion recognition systems, it is crucial to employ a variety of performance metrics to assess how well the models classify emotions and generalize across different datasets and scenarios. These metrics provide insights into the model's accuracy, robustness, and reliability. In particular, confusion matrices, ROC (Receiver Operating Characteristic) curves, accuracy, and loss graphs are some of the most commonly used tools to evaluate model performance in emotion recognition.

Confusion matrix

A complexity matrix is a table used to describe the performance of a classification model, usually by comparing predicted labels to the true labels from the test data. It provides a breakdown of correct and incorrect classifications for each class.

Each row of the matrix represents examples of true class labels, while each column represents examples of predicted class labels. The diagonal elements of the complexity matrix represent the number of correct predictions, while the off-diagonal elements indicate incorrect classifications, clearly showing where the model made mistakes.

ROC curve

The ROC curve is another valuable metric for evaluating the performance of emotion recognition models, especially in binary or multi-class classification problems. The ROC

curve plots the true positive rate (sensitivity) against the false positive rate (1 - specificity) at various threshold settings. A well-performing model will have a ROC curve that hugs the upper left corner of the graph, indicating a high true positive rate with a low false positive rate. The area under the ROC curve (AUC) is calculated to measure the performance of the model. An AUC score close to 1 indicates excellent performance, while a score close to 0.5 indicates that the model is performing no better than random chance.

Accuracy

Accuracy is perhaps the most straightforward and widely recognized performance metric in classification tasks, including emotion recognition. It is defined as the ratio of correctly predicted instances to the total number of instances. Although accuracy provides a general measure of the model's overall performance, it can sometimes be misleading in imbalanced datasets where certain emotion classes are much more frequent than others. Therefore, while accuracy is a useful metric, it is often supplemented by other metrics like precision, recall, and F1-score to provide a more complete picture of the model's performance.

Loss

The loss plot usually plots the training and validation loss over the training epochs. A well-performing model will exhibit decreasing loss for both the training and validation sets. If the validation loss starts to increase while the training loss continues to decrease, this is a sign that overfitting is occurring and regularization techniques or early stopping may be required. In short, it shows the relationships between the model's prediction and the actual data, which is a good guide for optimizing the model.

Results

Comparison of accuracy across models

In this section, the accuracy performance of each deep learning model used in the emotion recognition task is compared. Accuracy is a fundamental metric in classification tasks, providing a measure of how well a model can correctly predict the target labels across all classes. In this case, accuracy represents how effectively the models recognize various emotions from facial images.

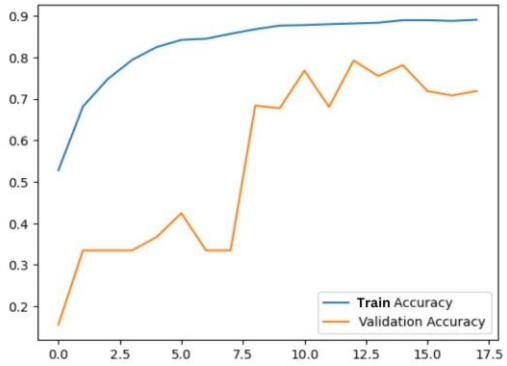


Figure 3. DenseNet-121 accuracy graphs.

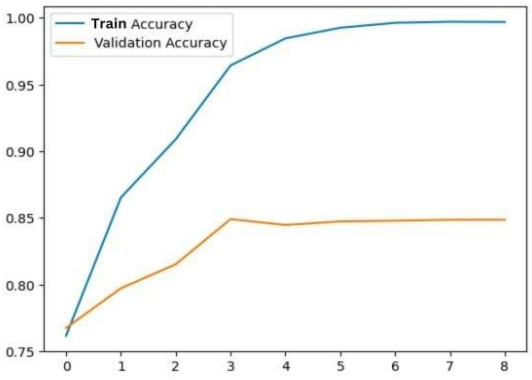


Figure 4. EfficientNet-B5 accuracy graphs.

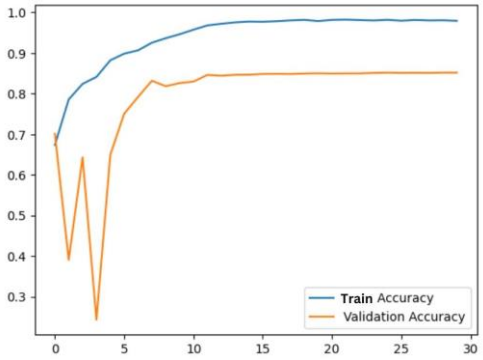


Figure 5. MobilenetV2 accuracy graphs.

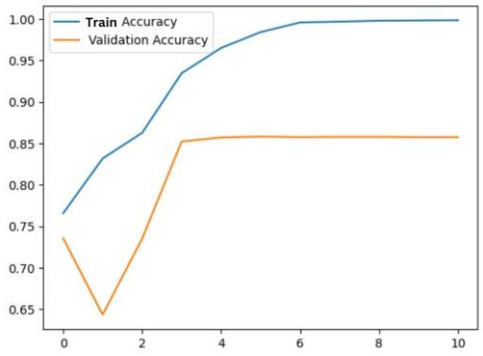


Figure 6. ResNet50 accuracy graphs.

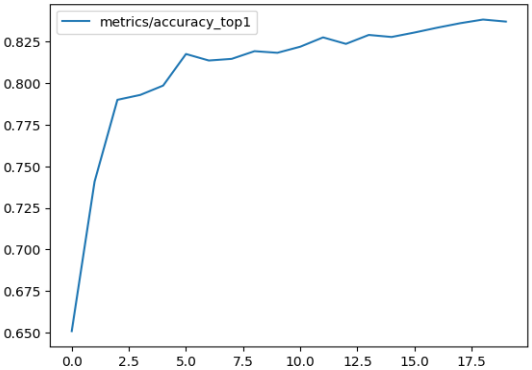


Figure 7. YOLOv8m-cls accuracy graphs.

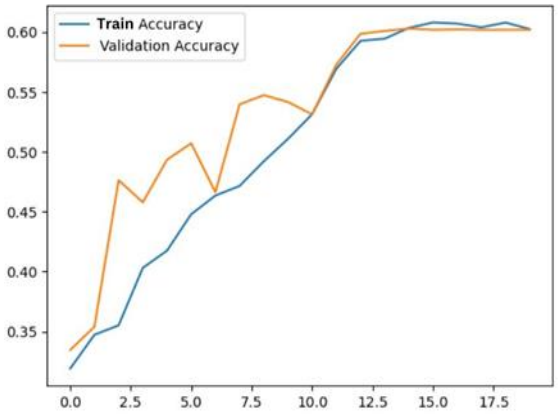


Figure 8. Vgg16 accuracy graphs.

Comparison of loss across models

Loss represents the difference between predicted and true values in classification tasks, and lower loss indicates better model performance. During the training process, models try to minimize the loss function, and monitoring both training loss and validation loss provides insights into potential problems such as overfitting or underfitting. Figures 9,10,11,12,13,14,15 show the losses.

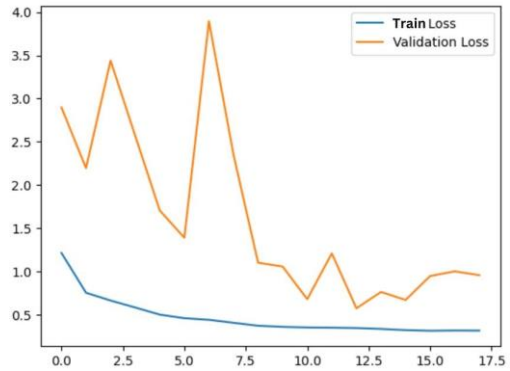


Figure 9. DenseNet121 loss graph.

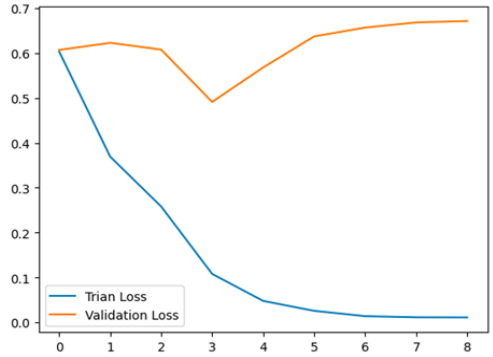


Figure 10. EfficientNet loss graph.

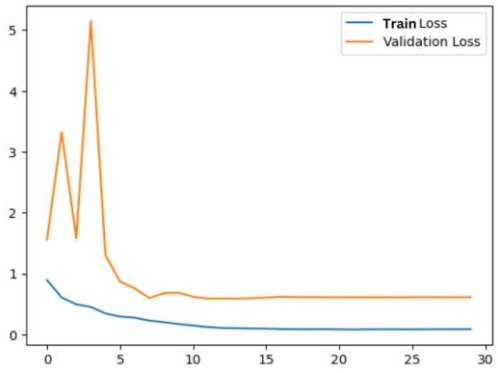


Figure 11. MobilNetV2 loss graph.

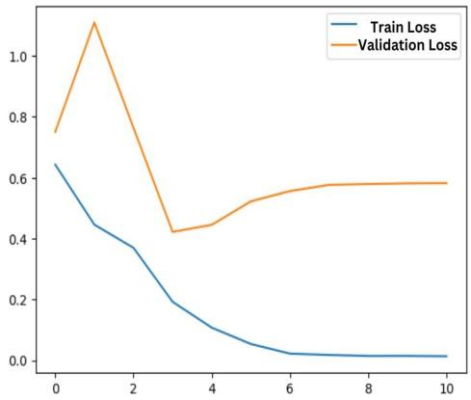


Figure 12. Resnet50 loss graph.

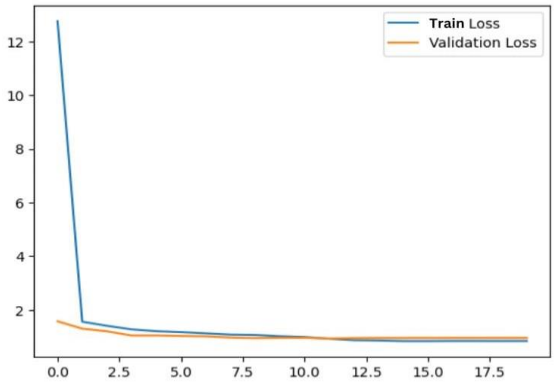


Figure 13. Vgg16 loss graph.

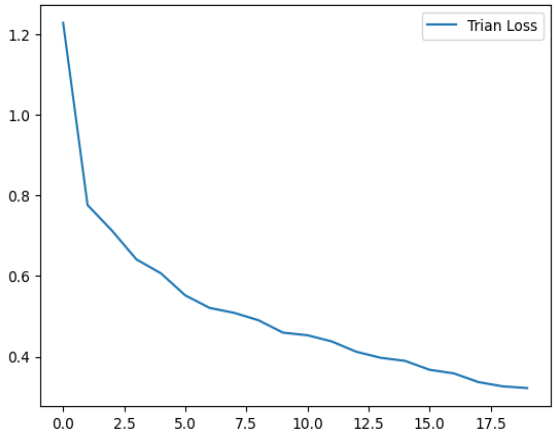


Figure 14. YoloV8m-cls loss graph.

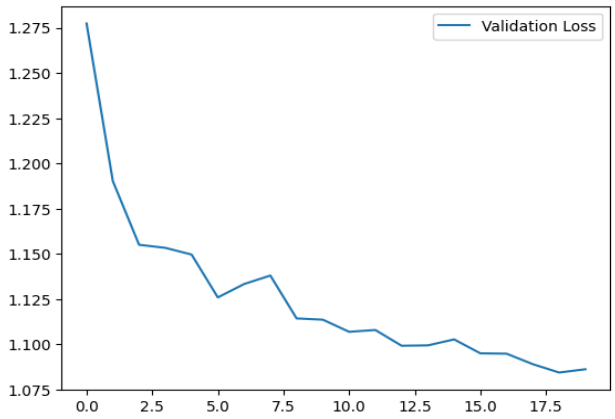


Figure 15. YoloV8-cls val/loss graph.

Table 3. Model loss values.

Model	Loss
DenseNet121	0,57
EfficientNet	0,49
MobilNetV2	0,61
ResNet50	0,52
VGG16	0,96
YoloV8m-cls	0,32

Confusion matrix comparison

The confusion matrix of the models is given in figures 16,17,18,19,20, and 21 below. By looking at this matrix, we can clearly see the relationship between the values predicted by the model and the actual values.

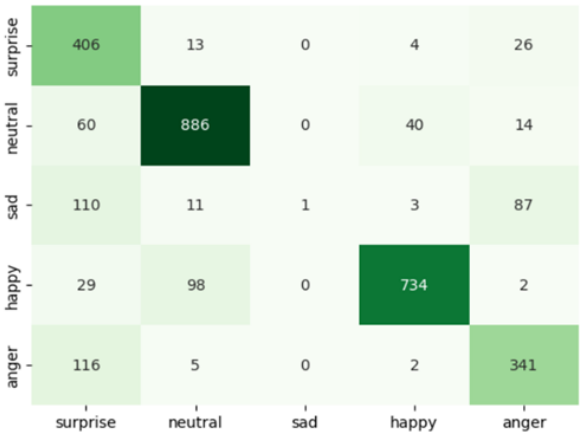


Figure 16. DenseNet121 confusion matrix.

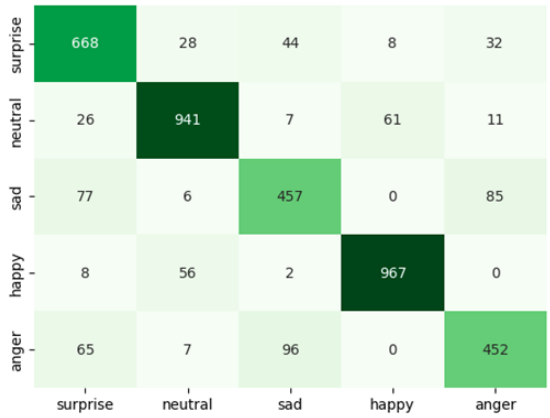


Figure 17. EfficientNet-B5 confusion matrix.

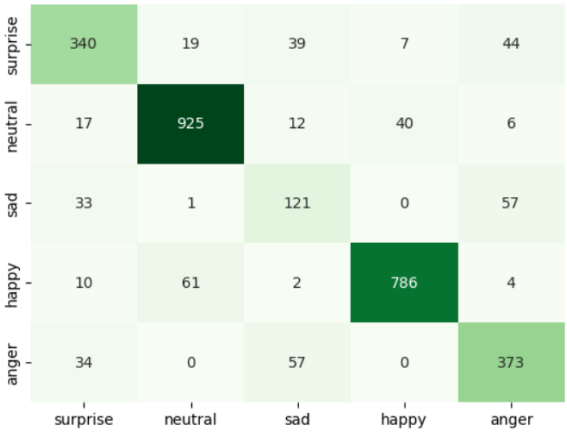


Figure 18. MobileNetV2 confusion matrix.

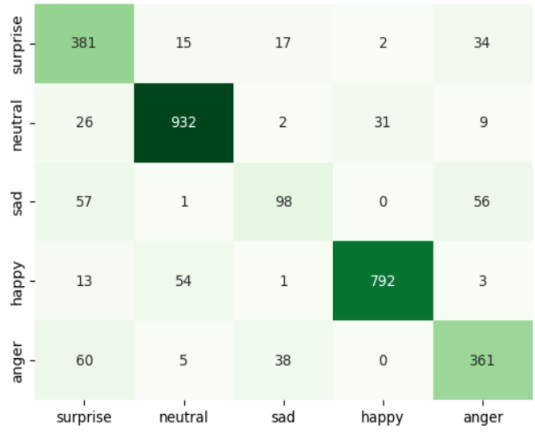


Figure 19. ResNet50 confusion matrix.

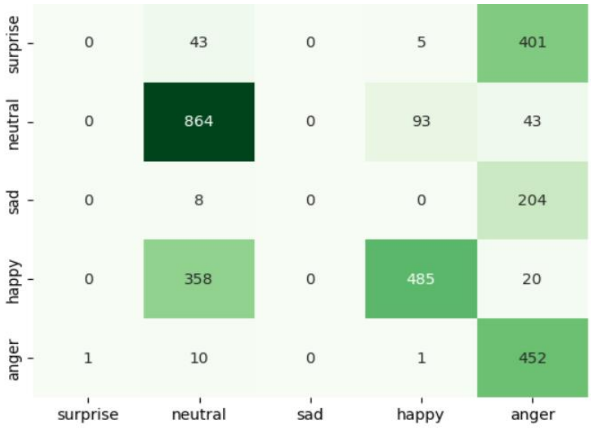


Figure 20. VGG16 confusion matrix.

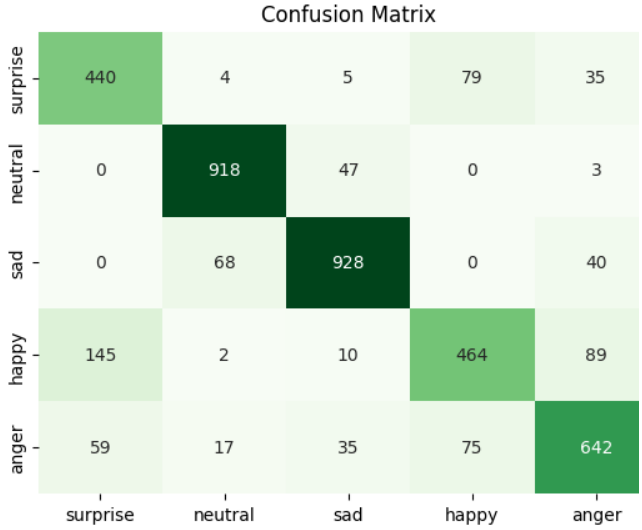


Figure 21. YoloV8m-cls confusion matrix.

The confusion matrix of each model is given in the outputs above. When the results are examined, it is seen that neutral and happy emotions are the best predicted classes in all models.

ROC curve comparison

ROC (Receiver Operating Characteristic) curve is another important metric to evaluate the performance of emotion recognition models. It graphically shows the trade-off between true positive rate (sensitivity) and false positive rate (1 - specificity) across various classification thresholds. Area under the ROC curve (AUC) is a widely used summary measure, where an AUC score closer to 1 indicates better performance. The Roc curve results for the models are given below in Figures 22, 23, 24, 25 and 26.

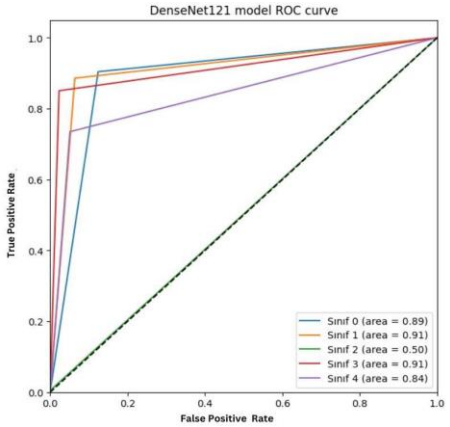


Figure 22. DenseNet121 roc curve.

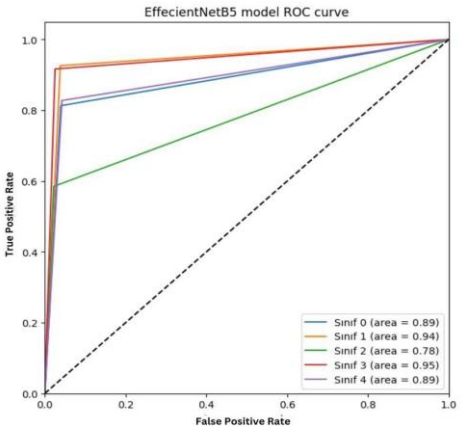


Figure 23. EfficienetNetB5 roc curve.

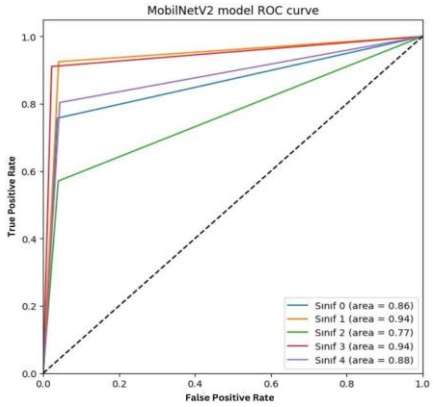


Figure 24. MobilNetV2 roc curve.

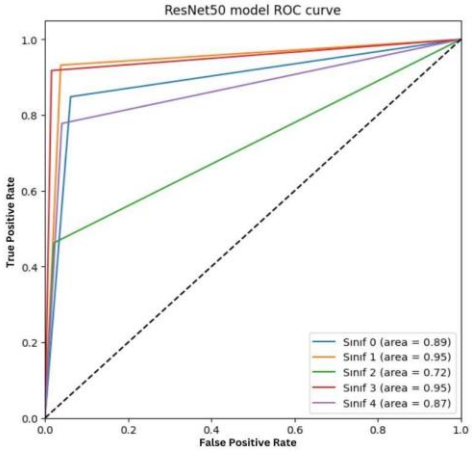


Figure 25. Resnet50 roc curve.

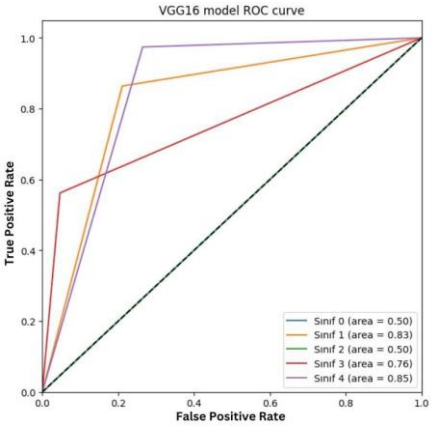


Figure 26. VGG16 roc curve.

The above curves show the Roc curves of all models, and while all models achieved a Roc curve above 0.85, only the VGG16 model achieved a score of around 0.75. The VGG16 model performed worse than the others in distinguishing emotions. It is easier to detect emotions of classes 1 and 3 (neutral and happy). YoloV8-cls was the best discriminator with 0.92.

Model Resource Consumption and Efficiency Analysis

Figure 27: Efficient Resource Consumption Chart

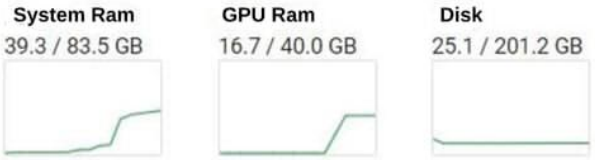


Figure 28. Resnet Resource Consumption Chart

Figure 29. Yolov Resource Consumption Chart

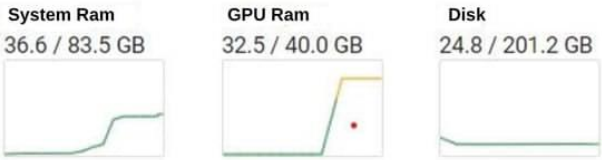


Figure 30. Desnet Resnet Resource Consumption Chart

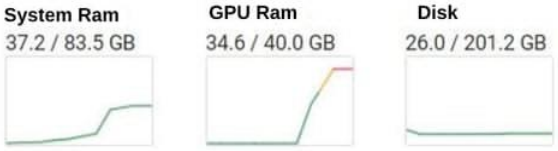


Figure 31. Vgg16 Resnet Resource Consumption Chart

Figure 32. MobilNet Resnet Resource Consumption Chart

Discussion

Table 4 below shows the metric results of all models used in the study. By looking at these results, we can compare the models with each other and form an idea about the loss values.

Table 4. Results of the models.

Model	Accuracy	Loss	Precision	Recall	F1-Score
DenseNet121	0,79	0,57	0,82	0,68	0,64
EfficientNet	0,84	0,49	0,82	0,81	0,82
MobilNetV2	0,85	0,61	0,79	0,79	0,79
ResNet50	0,85	0,52	0,8	0,79	0,79
VGG16	0,6	0,96	0,38	0,48	0,4
YoloV8m-cls	0,89	0,32	0,84	0,87	0,84

Analysis of best performing models

In this study, multiple deep learning models were evaluated for their performance on emotion detection tasks, and YOLOv8m-cls, ResNet50, and EfficientNetB5 emerged as the best performing models based on accuracy, loss values, and confusion matrix results. YOLOv8m-cls showed the highest overall performance, showing excellence in both accuracy and generalization. Its ability to maintain low loss values throughout the training process, combined with minimal misclassifications in the confusion matrix, made it the most reliable model for emotion detection. The strength of YOLOv8m-cls lies in its balance between speed and accuracy, which is crucial for real-time applications where fast detection is required. ResNet50 showed particularly strong results in terms of its ability to process complex facial features. It enabled the model to capture subtle patterns in facial expressions that may be lost in traditional convolutional layers. While ResNet50's training accuracy reached almost perfect levels, its validation accuracy faltered, indicating some overfitting. Despite this, ResNet50's overall performance, especially in distinguishing between subtle emotions such as anger and disgust, positioned it as one of the best performing models. With its efficient scaling of depth, width, and resolution, EfficientNetB5 offered a strong balance between accuracy and computational cost. While it did not outperform YOLOv8m-cls or ResNet50 in absolute terms, its composite scaling made it particularly valuable for scenarios where computational resources are limited. EfficientNetB5 showed a consistent reduction in both training and validation loss with minimal gap between the two, indicating good generalization and resistance to overfitting. MobileNetV2 showed solid performance in terms of speed and efficiency, but fell slightly behind the top three models in terms of accuracy. VGG16 struggled with overfitting, as evidenced by its high training accuracy but low validation accuracy. While VGG16's deep architecture allowed it to perform well on training data, it failed to generalize to unseen data, making it the weakest model in this study. The limitations of VGG16 suggest that it may not be suitable for emotion recognition tasks without further tuning, such as the addition of regularization techniques. YOLOv8m-cls stands out as the best overall performer, especially for real-time applications, while ResNet50 and EfficientNetB5 provide strong alternatives depending on the specific requirements of the task. The efficiency of MobileNetV2 makes it suitable for resource-constrained environments, but significant improvements are required for VGG16 to be applicable for emotion detection tasks.

Limitations of the models

While the deep learning models examined in this study show strong potential for emotion detection, they also exhibit several limitations that need to be addressed for real-world applications. Overfitting remains a persistent problem, especially for models such as VGG16 and DenseNet121. These models achieved high training accuracy but failed to generalize to validation and test sets, suggesting that they memorize training data rather than learn generalizable patterns. Computational cost is another limitation, especially for deep models such as ResNet50 and DenseNet121, which require significant processing power and memory to train and deploy. Although models such as MobileNetV2 and EfficientNetB5 are optimized for efficiency, they may sacrifice some accuracy in favor of lower computational demands. Limited training data is a major limitation. While large datasets such as AffectNet provide a solid foundation for training, emotion recognition datasets are still relatively small compared to datasets for other tasks such as object detection or natural language processing. The scarcity of labeled data makes it difficult for models to learn the full range of human emotional expressions, which can result in lower performance when used in a variety of real-world scenarios. The lack of multimodal data in this study is a limitation. Human emotions are often expressed through multiple channels, including facial expressions, voice, and body language. While this study focused only on facial expressions, future studies can integrate additional modalities to improve emotion recognition performance. Models that include audio or physiological signals, such as heart rate or skin conductance, can capture a more complete picture of emotional states and lead to more robust and accurate systems. Overfitting, computational cost, limited training data, and the lack of multimodal inputs are important hurdles that future research must address to develop more accurate, generalizable, and efficient emotion recognition systems.

Conclusion

Summary of findings

This study investigated the performance of various deep learning models (YOLOv8m-cls, ResNet50, EfficientNetB5, MobileNetV2, DenseNet121 and VGG16) on the task of emotion recognition from facial expressions. The aim was to identify the most effective model for emotion classification across 5 different emotion categories using transfer learning models. These emotions are Surprise, Neutral, Sad, Happy and Angry. Among the evaluated models, YOLOv8m-cls emerged as the best performing model overall. Its high accuracy, low loss values and strong performance in confusion matrix and ROC curve analysis indicated its suitability for real-time emotion detection tasks. ResNet50 and EfficientNetB5 also showed strong performance; ResNet50 showed excellence in capturing complex facial features and EfficientNetB5 balanced computational efficiency with

accuracy. Both models proved to be highly capable of generalizing to unseen data, making them viable options for a range of emotion recognition applications. Despite being slightly behind the top three models, MobileNetV2 showed commendable performance in terms of speed and efficiency, making it a good candidate for mobile or embedded systems where computational resources are limited. DenseNet121 struggled with overfitting issues despite its complex architecture, and VGG16 faced the most significant challenges, especially in generalization due to its tendency to memorize training data rather than learning generalizable patterns. The study also highlighted several key challenges in emotion recognition, including intra-class variability, inter-class similarity, environmental factors, and difficulty generalizing to new datasets. These challenges affected the performance of all models to varying degrees, and while YOLOv8m-cls and ResNet50 handled these challenges better than the others, no model was completely immune to misclassification or overfitting. It highlights the importance of choosing the right model architecture according to the specific needs of the application. It was found that YOLOv8m-cls was the most suitable solution for real-time applications, while ResNet50 and EfficientNetB5, which have different strengths in handling complex features and computational efficiency, offered strong alternatives.

Implications for future research

While the results of this study indicate significant advances in the use of deep learning models for emotion recognition, several areas remain ripe for future research. Addressing the limitations identified in this study, such as overfitting and computational cost, will be important to improve model performance and ensure scalability of emotion detection systems in real-world applications. Future work should focus on improving generalization capabilities.

Researchers should explore cross-dataset validation techniques where models are trained on one dataset and evaluated on another dataset to ensure that models can effectively generalize to unseen data. Another promising avenue for future research is the integration of multimodal data. Emotions are often expressed through a combination of facial expressions, audio, and physiological signals, and combining these modalities can significantly improve the accuracy and reliability of emotion recognition systems.

Multimodal models that include audio, text, or physiological data such as heart rate or galvanic skin response can provide a more comprehensive understanding of human emotions. Explainability and ethical concerns also represent important areas for future research. As emotion recognition systems are increasingly used in sensitive applications ranging from mental health to surveillance, the ethical implications of these technologies must be carefully considered. Ensuring privacy, bias reduction, and fairness in emotion recognition systems are important to prevent potential misuse or harm. Future research should focus on developing guidelines and frameworks to ensure these systems are used responsibly and equitably across all demographics.

Author contributions

AHI provided the idea for the software. Reviewed previous studies in this area and made suggestions for development with different methods and methods. MU completed the software part of the study and tested the results, compared them with each other and prepared a report. MFD provided planned progress in the writing phase of the article. Prepared the study according to the draft, taking into account the journal template and rules.

Acknowledgements

We would like to thank our valuable authors who made great efforts in the idea phase of this study, in the planning and software development of the article, and in transforming the results into an original article. We are also grateful to the journal members and referees who took the time to evaluate the article.

Declarations

Conflict of interest

The authors affirm that they have no known competing financial interests or personal relationships that might have influenced the work reported in this paper.

Consent to participate

All participants in this study provided informed consent to take part in the research.

Consent for publication

Authors have given their consent for the publication of the research findings and related materials

Financial support

No Financial Support was declared by the authors

References

[1] P. Ekman, "Facial expression and emotion," *Am. Psychol*, vol. 48, no. 4, pp. 384–392, 1993. DOI:10.1037/0003-066X.48.4.384

[2] M. Pantic and L. J. Rothkrantz, "An expert system for recognition of facial actions and their intensity," *AAAI/IAAI*, pp. 1026–1033, 2000.

[3] Z. Zeng, M. Pantic, G. I. Roisman and T. S. Huang, "A survey of affect recognition methods: Audio, visual and spontaneous expressions," *Proceedings of the 9th International Conference on Multimodal Interfaces*, pp. 126–133, 2007.

[4] J. Zhang, Z. Yin, P. Cheng and S. Nichele, "Emotion recognition using multi-modal data and machine learning techniques: A tutorial and review," *Inf Fusion*, 2020. DOI:10.1016/j.inffus.2020.01.011

[5] Y. Tian, T. Kanade and J. F. Cohn, "Facial expression recognition," in *Handbook of Face Recognition*, S. Z. Li and A. K. Jain, Eds., Springer, 2011, pp. 487–519. DOI:10.1007/978-0-85729-932-1_19

[6] P. Ekman and W. V. Friesen, *Facial Action Coding System*. Consulting Psychologists Press, 1978.

[7] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar and I. Matthews, "The extended cohn-kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression," *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition – Workshops (CVPRW),* pp. 94–101, 2010. DOI:10.1109/CVPRW.2010.
5543262

[8] G. Tolias and O. Chum, "Asymmetric feature maps with application to sketch-based retrieval," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3155–3164, 2017.

[9] A. Krizhevsky, I. Sutskever and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in Neural Information Processing Systems (NeurIPS),* pp. 1097–1105, 2012.

[10] D. Zilyas and A. Yılmaz, "Prediction model of educational success with machine learning methods," *DUMF Engineering Journal*, vol. 14, no. 3, pp. 437–447, 2023. DOI:10.24012/dumf.1322273

[11] I. Goodfellow, Y. Bengio and A. Courville, *Deep Learning*. MIT Press, 2016.

[12] M. Wöllmer, A. Metallinou, F. Eyben, B. Schuller, and S. Narayanan, "Context-sensitive multimodal emotion recognition from speech and facial expression using bidirectional LSTM modeling," *IEEE Transactions on Affective Computing*, vol. 2, no. 1, pp. 42–55, 2010.

[13] A. G. Howard *et al*., "MobileNets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint, arXiv:1704.04861*, 2017.

[14] F. Eyben *et al*., "The geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, 2015.

[15] J. Devlin, "BERT: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805,* 2018.

[16] A. Zadeh, M. Chen, S. Poria, E. Cambria and L. P. Morency, "Tensor fusion network for multimodal sentiment analysis," *arXiv preprint, arXiv:1707.07250*, 2017.

[17] R. T. Shawi, A. A. A. Abdulkadhim, and F. N. Abbas, "Facial Emotion Recognition Based on Improved ResNet50 Using Hybrid Pooling and Adaptive Leaky ReLU," *International Journal of Scientific Research in Science, Engineering and Technology*, vol. 12, no. 3, pp. 728–737, 2025.

[18] A. Alshammari and M. E. Alshammari, "Emotional facial expression detection using YOLOv8," *Engineering, Technology & Applied Science Research*, vol. 14, no. 5, pp. 16619–16623, 2024.

[19] A. A. Istiqomah, C. A. Sari, A. Susanto, and E. H. Rachmawanto, "Facial expression recognition using convolutional neural networks with transfer learning ResNet-50," *Journal of Applied Informatics and Computing*, vol. 8, no. 2, pp. 257–264, 2024.

[20] W. Du, "Facial emotion recognition based on improved ResNet," *Applied and Computational Engineering*, vol. 21, pp. 242–248, 2023.

[21] L. L. X. Wei and N. S. Sani, "Enhanced facial expression recognition based on ResNet50 with a convolutional block attention module," *International Journal of Advanced Computer Science & Applications*, vol. 16, no. 1, 2025.

[22] M. B. Sutar and A. Ambhaikar, "A Comparative Study on Deep Facial Expression Recognition," in *2023 7th International Conference on Intelligent Computing and Control Systems (ICICCS)*, pp. 903–911, May 2023.

[23] N. T. Singh, R. Ritu, C. Kaur, and A. Chaudhary, "Comparative analysis of traditional machine learning and deep learning techniques for facial expression recognition," in *2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, pp. 1–7, Jul. 2023.

[24] A. G. Vural, B. B. Cambazoğlu, P. Şenkul, and Z. O. Tokgöz, "A framework for sentiment analysis in Turkish: Application to polarity detection of movie reviews in Turkish," in *Computer and Information Sciences III,* E. Gelenbe and R. Lent, Eds., Springer, 2013, pp. 437–445. DOI:10.1007/978-1-4471-4594-3_42

[25] M. Bilgin and İ. F. Şentürk, "Sentiment analysis of document vectors based tweets using supervised and semi-supervised learning," *Journal of Balıkesir*

University Institute of Science, vol. 21, no. 2, pp. 822–839, 2019.

[26] G. M. Demirci, S. R. Keskin, and G. Doğan, "Sentiment analysis in Turkish with deep learning," *2019 IEEE International Conference on Big Data (Big Data).* DOI:10.1109/bigdata47090.2019.9006066

[27] M. Seyfioğlu and M. Demirezen, "A hierarchical approach for sentiment analysis and categorization of Turkish written customer relationship management data," *Proceedings of the 2017 Federated Conference on Computer Science and Information Systems (FedCSIS),* IEEE, 2017.

[28] B. Ayan, B. Kuyumcu and B. Ceylan, "Detection of Islamophobic tweets on Twitter using sentiment analysis," *Gazi Univ. J. Sci. Part C Des. Technol.*, vol. 7, no. 2, pp. 495–502, 2019.

[29] F. N. Uyaroğlu Akdeniz and H. I. Cebeci, "Sentiment analysis approach in evaluation of municipal services: Sakarya province example," *J. Intell. Syst. Theory Appl.*, vol. 4, no. 2, pp. 127–135, 2021.

[30] S. Tuzcu, "Classification of online user comments with sentiment analysis," *Eskişehir Turkish World App. Res. Center Informat. J.,* vol. 1, no. 2, pp. 1–5, 2020.

[31] D. Yohanes, J. S. Putra, K. Filbert, K. M. Suryaningrum, and H. A. Saputri, "Emotion detection in textual data using deep learning," *Procedia Comput. Sci.,* vol. 227, pp. 464-473, 2023.

[32] G. Zhang and S. Ananiadou, "Examining and mitigating gender bias in text emotion detection task," *Neurocomputing*, vol. 493, pp. 422–434, 2022. DOI:10.1016/j.neucom.2022.03.051

[33] D. Issa, M. F. Demirci, and A. Yazici, "Speech emotion recognition with deep convolutional neural networks," *Biomedical Signal Processing and Control*, vol. 59, p. 101894, 2020. DOI:10.1016/j.bspc.2020.101894

[34] A. S. Yuksel and F. G. Tan, "Knowledge discovery in social networks with text mining techniques," *J. Eng. Sci. Design*, vol. 6, no. 2, pp. 324–333, 2018. DOI:10.21923/jesd.384791

[35] M. Turan and E. Arığ, "Video sentiment analysis," *Eur. J. Sci. Technol. (EJOSAT) Supplementary Special Issue (HORA),* pp. 34–41, 2021.

[36] C. D. Eyipınar, F. Büyükkalkan and K. Semiz, "Sentiment analysis of YouTube video comments on athlete nutrition," *Int J Phys Educ. Sports Technol*., vol. 2, no. 2, pp. 27–39, 2021.

[37] A. Tepecik and E. Demir, "Analysis of Turkish voice recording data labeled with three emotions using machine learning algorithms," *Gazi Univ. Fac. Eng. Arch. J.,* 39(2):709–716.

[38] A. Sel, "Analysis of public opinion during the pandemic period using sentiment analysis method: The case of Türkiye," *Beykoz Academic Journal*, vol. 10, no. 2, pp. 134–154, 2022.

[39] M. Uysal, "Comparison and applicability of modifying deep learning models for emotion recognition," M.S. thesis, Burdur Mehmet Akif Ersoy Univ., 2024.

[40] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L. C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* pp. 4510–4520, 2018.

[41] K. He, X. Zhang, S. Ren and J. Sun, "Deep residual learning for image recognition," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* pp. 770–778, 2016.

[42] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556,* 2014.

[43] J. Redmon, S. Divvala, R. Girshick and A. Farhadi, "You only look once: Unified, real-time object detection," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* pp. 779–788, 2016.