Article
JOURNAL OF UNIVERSITY OF BABYLON
For Pure and Applied Sciences (JUBPAS)
Vol. 34 ; No.1 | 2026

A Hybrid Transfer Learning Framework for Efficient Facial Keypoint Detection

Shahad Eaad Abdulmahdi[1*]
[1]Islamic University of Lebanon (IUL), shahad.eaad95@gmail.com, Lebanon

***Corresponding author email:** shahad.eaad95@gmail.com; mobile: 07811959595

إطار عمل هجين للتعلم الانتقالي من أجل الكشف الفعال عن النقاط الرئيسية للوجه

شهد ايعاد عبدالمهدي [1] *
الجامعة الإسلامية، shahad.eaad95@gmail.com ، بيروت ، لبنان

Accepted: 1/3/2026 **Published:** 31/3/2026

ABSTRACT

Background:

Face keypoint detection is one of the basic supporting tasks for applications ranging from emotion and facial expression human-computer interaction. Typical deep learning models struggle to make a trade-off among accuracy, generalization, and computational efficiency.

Materials and Methods:

A hybrid deep learning model is proposed in this paper by stacking three pre-trained CNNs: VGG16, ResNet50, and MobileNetV2 to improve both accuracy and speed of facial landmark localization. This contains several steps of data preprocessing-normalization, converting into grayscale, resizing images to 150 × 150 pixels; using MTCNN for face detection and ROI extraction. In the extraction phase, fully connected layers predict the 2D coordinates of 15 key facial landmarks based on concatenated output features from three parallel CNN branches. The network is trained with the optimizer [learning rate = 0.001], Mean Squared Error(MSE) as the loss function over 50 epochs.

Results:

The hybrid model showed good convergence and stability with final error values of MAE = 3.3, MAPE = 7.6, and RMSE = 4.39. From both quantitative and qualitative analysis results, the predicted keypoints nearly match ground-truth values for different facial images clearly proving the model to be robust as well as having high localization accuracy.

Conclusion:

The proposed hybrid architecture effectively combines several CNN models to optimize and balance representational depth, feature diversity, and computational efficiency. It attains high performance and generalizability in facial keypoint detection, thus making it a potential solution for state-of-the-art applications in emotion recognition, facial analytics, and human-computer interaction

Key words: Facial Keypoint Detection; Deep Learning; Hybrid Model; Convolutional Neural Networks; Computer Vision..

مجلة جامعة بابل للعلوم الصرفة والتطبيقية مجلة جامعة بابل للعلوم الصرفة والتطبيقية
ISSN: 2312-8135 | Print ISSN: 1992-0652
info@journalofbabylon.com | jub@itnet.uobabylon.edu.iq | www.journalofbabylon.com

ISSN: 2312-8135 | Print ISSN: 1992-0652
info@journalofbabylon.com | jub@itnet.uobabylon.edu.iq | www.journalofbabylon.com

INTRODUCTION

Facial keypoint detection (aka, landmark localization) is an important task in computer vision, with numerous downstream applications such as facial recognition, expression recognition, augmented reality, driver monitoring systems, and more [1], [2]. The task in question is to accurately predict a set of coordinates (e.g. the corners of eyes, edges of mouths, tip of nose, location of eyebrows) on an image of a face, allowing for rich modeling of the structure, motion, and semantics of the face.

Traditional solutions to this challenge typically involve either hand-crafted features and geometric models, such as Active Shape Models (ASM), Active Appearance Models (AAM), or feature descriptors like SIFT, SURF, and ORB [3] – [7]. Whilst working reliably in controlled conditions, such methods tend to have limited robustness to large pose variation, occlusion, illumination change, and unconstrained real-world images [8]–[10]. Classical registration or alignment pipelines (intensity-based or feature based) may rely on some careful initialization and underperform on datasets with varied facial conditions [11], [12].

Unlike previous methods, deep learning has made dramatic advances in the ability of computers to automatically learn to detect unique positions (landmarks) on faces. CNNs are particularly good at modeling a picture of a face and using both small parts and large parts of the face to produce its unique landmarks, even when the face is distorted in a variety of ways [13]–[15]. However, creating a new model using a deep learning architecture to detect keypoints on a face is not easy; (1) there are not enough available datasets that have many images of faces annotated with their respective facial landmarks; (2) deep learning architectures require significantly more computing resources to train; and (3) there is strong emphasis that the new model can generalize across many different conditions that each are unique and vary from one another.

Transfer learning is one of several strategies used to address the aforementioned limitations. Deep networks that were pre-trained using large datasets (e.g., ImageNet) can be fine-tuned and thus accurately adapted to the target application (i.e., a specific dataset). Because of this, transfer learning utilizes low- and mid-level visual representations (e.g., edges, textures, shapes) to affect model convergence time and also helps with reducing overfitting, increasing generalization ability; all within a small dataset size [16]–[18]. Transfer learning also enables the sharing of the same architecture (i.e., VGG, ResNet, MobileNet) by facial keypoint detection models that learned rich and generalizable input features through previous tasks.

This paper provides an overview of a hybrid deep learning architecture designed for facial landmark detection by combining two separate networks, VGG16, ResNet50, and MobileNetV2, for a single framework. The model operates on a set of preprocessed, grayscale images resized to 150 x 150 pixels after applying facial cropping with MTCNN, as well as on the combined output of each network's features. VGG16 captures fine detail (textured and edge) while ResNet50 provides a greater level of hierarchical abstraction because of its use of residual connections compared to VGG16. Likewise, MobileNetV2 has advantages over VGG16 and ResNet50 with respect to computational performance and uses depth-wise separable convolutional layers to provide efficiency. All three types of features/representations are

ISSN: 2312-8135 | Print ISSN: 1992-0652
info@journalofbabylon.com | jub@itnet.uobabylon.edu.iq | www.journalofbabylon.com

combined and normalized before processing through fully connected layers resulting in a total of 30 outputs that each correspond to 15 (x,y) landmark produced from a face detection process[19].

Adam Optimizer with Mean Squared Error loss (MSE) and Dropout for Overfitting [20] - [22]. Performance was measured on a basis of Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) located between predicted versus actual [23], [24]. The Hybrid Scheme strives for balanced accuracy of models versus speed of inferences versus generalization across unconstrained facial domains.

The remainder of this document is structured as follows: Section II explains the datasets used for this study and their corresponding processing pipelines. Section III discusses the architecture hybridization used to create the hybrid architecture, including training protocols. Section IV provides information on experimental findings using the hybrid architecture and comparison methods; Section V discusses the applications/limitations of this research; and, finally, Section VI summarizes this research and provides directions for future studies.

MATERIALS AND METHODS

Several studies worked on diagnosing and analyzing brain tumor images via several intelligent artificial intelligence program designs and implementations.

1. Detecting facial keypoints, such as eye corners, eyebrow tips, the tip of the nose, and end points of the lips, are some of the most difficult tasks in computer vision due to the vast amount of data used to train a machine learning model on the location of facial keypoints on a given image. This paper presents an attempt to complete this difficult computer vision task through the use of a Convolutional Neural Network (CNN) to predict where facial keypoints will appear in webcam impages. Through the use of "truth" keypoints on individual images taken from the webcam, we were able to predict the location of keypoints using each CNN model, and therefore have been able to measure how accurately each model predicts the locations of keypoints. To further improve the accuracy of each model, we were able to create a numerical measure of how much error the models produce by comparing the predicted positions of keypoints with the truth keypoints from collected webcam images using a Mean Squared Error (MSE) metric. This method combined with the deep learning capabilities of each CNN produces promising results for improving the accuracy of facial feature localisation and the overall growth of computer vision in other similar tasks[25].
2. In this study, we developed a method to determine facial landmarks based on a deep learning framework to support keypoint detection from dynamic media. Because social media utilizes both augmented reality and image filters, delivering sufficient accuracy requires a precise and repeatable method for estimating the location of facial landmarks. To achieve this, a Convolutional Neural Network (CNN) based upon MobileNetV2 was used as it has demonstrated superior performance to other traditional approaches. Each of the 15 key points is estimated using a two-dimensional coordinate system representing a minimal non-redundant description of the human face. A reference model (model_rit_lhn) was built first with a single hidden layer neural network and trained. Then, a new CNN-based upon MobileNetV2 was generated, and applied to facial images captured in different

ISSN: 2312-8135 | Print ISSN: 1992-0652
info@journalofbabylon.com | jub@itnet.uobabylon.edu.iq | www.journalofbabylon.com

conditions of pose and lighting, which would typically hinder facial landmark localization accuracy. The model was rated using the code based on an archived dataset and completed with a difficult problem of facial keypoint detection system. The accuracy of each model was validated at 84%, where the proposed MobileNetV2 architecture produced the greatest results compared to other modern day models. The above design indicates how well a lightweight deep learning model can be used to enhance both the performance and the reliability of a facial keypoint detection system[26].

3. This study uses a machine learning approach to combine facial point detection with emotion classification based on facial expressions to improve the accuracy of determining an individual's emotions. Facial expressions convey emotions and serve as a key way of non-verbal communication. They can vary in their intensity from one area of the face to another, especially the mouth and eyes, so the research framework will focus on identifying and analyzing those areas of the face that demonstrate the greatest amount of expression. The methodology begins with locating the facial and extracting meaningful key points from that face; these points will then be used as inputs to encode the facial expression by gesture recognition systems. Following this, a hybrid approach will be developed using neural networks to extract features and construct classification algorithms to classify emotion from the various expressions. The two-phase process will ensure that both the features used to provide facial landmarks and the emotive classification will be robust to both lighting and camera angle. Using accurate performance measurements and benchmark test datasets to perform validation for evaluating the success of combining keypoint detection and expression analysis in real-world situations (health monitoring, surveillance, human-computer interaction) [27].
4. The motivation of the proposed method for facial paralysis detection is to extract and analyze Facial landmarks for quantifying asymmetry between the two sides of a face. As shown in Fig. 1a, the proposed approach starts from an input frontal face image that is firstly scaled to a same size by gray and scale processing. At these cropped regions, the face is detected using publicly available dlib libraries and MEE shape predictor is employed to obtain 68 facial landmarks. Among these, 51 points were considered the most relevant for facial measures computation . In addition, to remove the effect of head tilt on accuracy, the angle of rotation of face was calculated and adjusted using jaw points 48 and 49 for feature extraction.

   Quantitative facial measurements were then derived according to spatial distances between the extracted set of landmarks including eyesbrows, eyes, nose and mouth. In all, 34 distances were obtained (A–Z) (Fig. 3), from A to Q to evaluate structural asymmetry between right and left sides of the face, whereas R to Z considered facial movements. These extracted myriad features were then used as an input to a Multi-Layer Perceptron (MLP) that performed classification on the input as healthy or impaired, thus forming the classification pipeline[28].
5. The proposed head pose estimation system is organized into two main stages: 3D face reconstruction and 3D–2D keypoint matching. During the reconstruction process, a personalised 3D face model is generated from a RGB input using convolutional neural networks (ResNet-101 and ResNet-18) designed to be optimised together with an asymmetric Euclidean loss and facial feature points (FFP) loss ensuring good geometric and expression accuracy. The high dimensionality of the mesh representation is reduced using PCA, so that both identity-sensitive geometry and facial expressions can be captured.

ISSN: 2312-8135 | Print ISSN: 1992-0652
info@journalofbabylon.com | jub@itnet.uobabylon.edu.iq | www.journalofbabylon.com

In the second stage, a coordinate descending iteration algorithm is adopted to achieve the reconstructed 3D facial keypoints matching with 2D detected landmarks under weak perspective transformation and obtain yaw, pitch,and roll angles effectively.
Extensive experimental results on five benchmark databases (Pointing'04, BIWI, AFLW2000-3D, Multi-PIE and Pandora) prove the robustness and effectiveness of our method. Results show that the FFP loss is able to incorporate lost details thereby improving the reconstruction quality in terms of lower MAE values over those methods trained with standard Euclidean loss. Besides, the proposed method outperformed or was comparable to the state-of-the-art on different datasets in average MAEs from 5.81° to 7.59° with various keypoint detector settings and retained robustness during occlusion conditions. Crucially, the proposed model demonstrated excellent cross-dataset generalization performance compared to the state-of- the-art supervised methods that are susceptible to label bias. These results demonstrate that the combination of tuned 3D reconstruction with iterative 3D–2D keypoint matching offers a robust and effective solution for head pose estimation[29].

6. The proposed FER method is designed based on an occluded expression recognition model that is developed under the framework of generative adversarial network (GAN). The system consists of two major modules: (1) occluded face image restoration to reconstruct the incomplete or masked facial region to retrieve discriminative expression features; and (2) facial expression recognition, where a deep CNN is used to classify the restored images into one of several expression categories. In order to lay the research foundation, this paper first investigates the deep learning–based facial expression recognition approaches in both static and dynamic classifications during ten years by reviewing and analyzing them. Also, several benchmark datasets widely utilized for facial expression analysis are surveyed to enhance the assessment of algorithm performance. We also compare the state-of-the-art methods through experiments to further demonstrate the benefits and drawbacks of these existing models under real environment, taking account factors like data scarcity resulting in over-fitting and disturbance from non-expressive facial movements[30].
7. We introduce a novel method referred to as Keypoint-based Relative Position Encoding (KP-RPE), which incorporate facial keypoints into the relative position encoding scheme of ViTs for recognition tasks on unconstrained and low-quality images. In contrast to the traditional face recognition methods which only understand well-aligned facial inputs, KP-RPE takes advantage of encoding the spatial information among image patches and key facial landmarks (such as eyes, nose and mouth) in an explicit attention mechanism.The approach is based on the standard self-attention of Transformer models, which splits each image into non-overlapping patches and tokenizes them. We propose a keypoint-dependent bias matrix to extend traditional relative position encoding, which modifies the weights of attention according to the relative spatial distance between query–key pairs and their association with facial landmarks detected. To achieve this, facial landmarks are located by state-of-the-art detectors like MTCNN or RetinaFace and projected to the image grid (normalized). Three variants of KP-RPE are formulated: (1) absolute keypoint encoding, which encodes normalized keypoint coordinates to offset directly;(2) relative keypoint encoding, which computes the offsets with distances regarding keypoints and image patch; and (3) multi-head relative encoding further makes the query–keypoint interaction differentiable across various self-attention heads.This cross-layer fusion enables ViTs to dynamically refine attention maps according to facial structure cues, making them robust

ISSN: 2312-8135 | Print ISSN: 1992-0652
info@journalofbabylon.com | jub@itnet.uobabylon.edu.iq | www.journalofbabylon.com

over image mislignment, low resolution and occlusion. As shown in Figure 4 of the original paper, KP-RPE can be easily incorporated into each multi-head self-attention block with almost insignificant computation costs, which demonstrates its scalability and effectiveness for facial tasks in diverse real-world environments[31].

8. The facial recognition model in our experiment has been created using a geometric neural network, which uses point clouds to extract and identify facial expressions. To test our model's performance, we used two benchmark datasets. The first was the Bosphorus Database, which contains data from 65 individual participants identifying 7 different types of basic facial expressions (anger, disgust, fear, happiness, sadness, surprise and neutral). The other dataset used by our experiment was the SIAT-3DFE data set, which contains facial expression recognition data from 150 individual participants identifying 4 different types of basic facial expressions (neutral, happiness, sadness and surprise). In order to produce high quality the final dataset that we used to train our models, we performed preprocessing operations designed to produce more diverse, higher quality samples of the data. These preprocessing operations included performing methods such as face-centre cropping, augmenting the training data, and densifying the point clouds by denoising the point cloud. The extracted 3D shape features from each of the point cloud scans were then extracted and classified using the PointNet++ model. We further improved the performance of our model through the use of hyperparameter tuning and cross-validation testing. Finally, we evaluated how well our system performed during the evaluation phase of the study using recognition accuracy and confusion matrices across both datasets. According to research results, a 69% accuracy rate can be identified as achieved with seven expression categories of Bosphorus database and by reducing them down to five (anger; disgust; happiness; surprise; neutral), an increase in success to an 85%. The model yielded an 78% success rate on the SIAT-3DFE standard. This indicates that models utilizing geometric deep learning to analyze 3D point cloud representations of faces can successfully provide solutions against challenges associated with recognizing the positions of human faces, and could possibly be used within rehabilitation programs for individuals suffering from facial palsy [32].
9. The proposed facial key-point detection research methodology has three main stages. " "Augment the data augmentation on the training dataset, we used a variety of transformations to increase the number of samples and improve model generalization. Flicker Strobing There were augmentation methods that consisted of geometric and photometric transforms, generating various transformed samples while maintaining the facial structures of its subjects. In the second stage, an Inception-based deep neural network structure was designed to detect facial key-points effectively. The model was constructed with the aim of minimizing training duration without sacrificing accuracy in identifying key regions on frontal facial images, such as the eyes, eyebrows, nose and lips. Finally, at testing a test-time average of the predictions over multiple augmentations of the same image guaranteed good performance and robustness.The performance of the proposed model in comparison with and without augmentation was quantified as MSE. Experimental results showed that our model dramatically reduced the MSE of facial key-point detection in compared with current state-of-the-arts one, suggesting effectiveness to be used for accurate and efficient facial key-point detection[33].
10. The novel approach to the driver fatigue detection in this paper mainly focuses on improving accuracy and efficiency for smart cars. The first step is facial key-point

ISSN: 2312-8135 | Print ISSN: 1992-0652
info@journalofbabylon.com | jub@itnet.uobabylon.edu.iq | www.journalofbabylon.com

detection for which a model is trained using multi-block local binary patterns (MB-LBP) and Adaboost classifier to detect 24 key-points of the face. From these landmarks we detect the states of eyes and mouth to recognize fatigue behaviors. Two fatigue measures are calculated:the percentage of eye closure over time (PERCLOS), and the frequency of yawning, both being quantitative indicators of fatigue. In the last step, a Fuzzy system is used to categorize the driver state into three different levels: normal, fatigued or highly fatigued. Experimental results show that the proposed method has excellent detection precision and sensitivity to fatigue status, in consideration of its rapid responses, reliability, high accuracy and therefore it is practical and effective for real-time eye-fatigue monitoring in intelligent en-vehicle system[34].table 1 shows summary for lecture views .

Table 1 : the summary of lecture views

Study	Technique Used	Main Goal / Results	Key Limitations
[25] Colaco & Han (2020)	CNN + MSE	Detect facial keypoints (eyes, nose, mouth) from webcam images as a baseline model	Simple design, limited dataset, weak generalization
[26] Kulkarni et al. (2021)	MobileNetV2 (lightweight CNN)	Detect 15 facial points with 84% accuracy and high computational efficiency	Limited to a small set of points, reduced expressiveness
[27] Sooch & Anand (2021)	Hybrid model (CNN + classification algorithms)	Combine keypoint detection with emotion classification for improved accuracy	Sensitive to changes in lighting and pose
[28] Parra-Dominguez et al. (2021)	dlib + MLP	Diagnose facial paralysis via asymmetry analysis using 34 spatial distances	Relies only on frontal images, limited diagnostic scope
[29] Liu et al. (2021)	CNN (ResNet-101/18) + 3D Reconstruction	Estimate head pose (Yaw, Pitch, Roll) accurately across multiple datasets	High computational cost, longer processing time
[30] Ge et al. (2022)	GAN + CNN	Restore occluded facial regions and classify expressions with high accuracy	Risk of overfitting due to limited data
[31] Kim et al. (2024)	Transformer (ViT) + KP-RPE	Integrate keypoints into attention mechanism for robust recognition on low-quality or misaligned images	Complex implementation, requires large computational resources

ISSN: 2312-8135 | Print ISSN: 1992-0652
info@journalofbabylon.com | jub@itnet.uobabylon.edu.iq | www.journalofbabylon.com

[32] Nguyen et al. (2021)	PointNet++ (Geometric Deep Learning)	Use 3D point clouds for facial expression recognition, achieving up to 85% accuracy	Lack of diverse and large-scale 3D datasets
[33] Dwivedi & Sharan (2022)	Inception CNN + Data Augmentation	Reduce MSE and enhance generalization in facial keypoint detection	Tested only on frontal images, limited real-world evaluation
[34] Liu et al. (2020)	MB-LBP + Adaboost + Fuzzy System	Detect driver fatigue (eye closure, yawning) with high accuracy and speed	Limited generalization across ethnicities and lighting conditions in vehicles

In table1 recent advances in facial keypoint detection Research on facial keypoint detection has made significant strides both in methodologies and applications. From earlier studies which utilized CNN's for the first time [25] to locate main facial landmarks, to now. These early methods were already being utilized with various advanced techniques such as Inception Architectures and Data Augmentation. The result was improved performance and reduced error rates. Subsequent work applied more advanced models has further increased efficiency and reduced error rates [33]. In addition to these studies, many additional studies were conducted to achieve high performance using very lightweight models (e.g MobilNetV2 [26]) or using transformer models with Keypoint Relative Positional Encoding (KP-RPE) [31] providing a means to ensure the robustness of the recognition system under low-quality and/or misaligned images. Further examples of this include utilizing facial keypoint detection as a foundation for complex applications; such as determining emotion and expression classifications by augmenting with additional modules [27] and/or the ability to handle occlusion using generative adversarial networks (GANs) [30] 3D data and geometric deep learning (PointNet++) have introduced advances that improved performance and made it easier to apply machine learning in clinical settings. Direct examples of using these technologies in medicine are facial paralysis diagnosis through asymmetry analysis and pose estimation of 3D reconstructed images. Practical applications of these technologies include developing intelligent systems that detect driver fatigue based on eye and mouth characteristics using fast algorithmic methods such as Adaboost and MB-LBP. In general, research has moved steadily from the establishment of benchmark models for the detection of landmarks toward improving the accuracy and efficiency of such models and integrating those improvements into real-world applications, including health care, security, and end-user systems, through the implementation of the latest artificial intelligence technologies to provide reliable results for the applications of this work. The summary of related work is itemized in Table 1.

ISSN: 2312-8135 | Print ISSN: 1992-0652
info@journalofbabylon.com | jub@itnet.uobabylon.edu.iq | www.journalofbabylon.com

Transfer Learning models

As illustrated in Figure 1, transfer learning has emerged as one of the most prevalent paradigms for leveraging deep neural networks to solve many of today's most important machine learning problems. In general, transfer learning involves using a pre-trained model, along with an associated task set of data, to adapt that model to other related (or similar) domains. This is accomplished by taking advantage of the deep neural networks learned weights and hierarchical features, thereby eliminating the need for computationally expensive training processes from scratch.

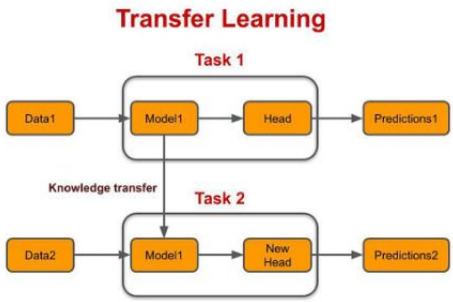


Figure 1: Transfer machine learning [30].

The most significant idea behind retraining only part of the network while keeping the rest of the already trained network the same as before, is that you can take advantage of previously learned information again when you have a second training set which is related in some way to your first training set but much smaller in size. The result is large gains in training efficiency, convergence time and general accuracy when time/resources are limited, or there is little labelled data available [30]. Numerous major pre-trained models that have been created across different domains — e.g. image classification, natural language processing, temporal (time) trend prediction — have strong network architectures that help them generalise extremely well; for example, the DenseNet (Densely Connected Convolutional Networks) [31–32] architecture which addresses issues with the vanishing gradient problem by creating direct connections between all the layers of the network ensuring that maximum amount of information can flow through to the final layer and reuse features throughout the network. In contrast to deep networks that increase their width/depth, DenseNets primarily focus on connectivity. Both models of DenseNet - DenseNet121 and DenseNet201 - can be found in the literature indicating the depth at which these designs operate. Another architectural model; MobileNet V3, which is an offshoot of Google's MobileNet, focuses entirely on mobile and embedded devices. MobileNet V3 combines a lightweight design with high levels of accuracy while maintaining low latency. MobileNet V3 includes additional features over MobileNet V2, including inverted residual blocks, squeeze-and-excitation (SE) modules, and a new hard-swish activation function. Two versions of MobileNet V3 have been developed through neural architecture search (Nes), one intended for low-end resource environments and another intended for high-performance applications: MobileNet V3 small and MobileNet V3 large [33]. They offer an effective and scalable solution via transfer learning to various medical image analysis problems, including the detection of skin cancer where accurate and timely detection is critical.

ISSN: 2312-8135 | Print ISSN: 1992-0652
info@journalofbabylon.com | jub@itnet.uobabylon.edu.iq | www.journalofbabylon.com

MobileNetV3 Algorithm

The Google team created MobileNetV3, the most recent iteration of the MobileNet series, to offer a deep model with excellent accuracy and execution speed while consuming minimal resources, which makes it perfect for embedded systems and mobile devices. In addition to incorporating powerful components like MobileNetV2 inverted residuals, SE (Squeeze-and-Excitation), and a new activation called Hard-Swish, a more effective alternative to Swish, MobileNetV3 combines a number of architectural improvements, building on the outcomes of automatic search using AutoML techniques (specifically, NAS - Neural Architecture Search). There are two iterations of the model: MobileNetV3-Small for devices with modest capabilities and MobileNetV3-Large for excellent performance **[33].**

ShuffleNet-V2

ShuffleNet v2 was offered as a direct improvement over ShuffleNet v1 after researchers realized that various theoretical notions employed in lightweight network design do not necessarily convert to actual performance on real-world devices such as smartphones. The paper outlines four guidelines for designing more efficient lightweight networks: minimizing the number of pointwise convolutions due to their high resource consumption; minimizing memory access overhead (MAC) due to its greater impact on model speed than the number of FLOPs; avoiding excessive channel grouping, which has a negative impact on channel communication; and minimizing non-parallel operations, such as summation in short paths, due to their slowdown**[34]**.

Evaluation Metrices

The evaluation of how well a machine learning model works is a basic component in proving its efficiency most especially classification models. There are some standard metrics that are commonly used to determine the ability of a model to predict. Among these include Accuracy which is simply the proportion of correctly classified samples to the total number of predictions as expressed in equation(1).

$$\boldsymbol{accuracy} = \frac{\boldsymbol{number\ of\ correct\ prediction}}{\boldsymbol{total\ number\ op\ prediction}} \qquad (1)$$

But accuracy can never be enough, especially in the case of unbalanced datasets. To add more perspective, a confusion matrix is mostly used which gives a detailed account of the results of predictions in terms of true positives, false positives, true negatives and false negatives as shown in Figure 2.

ISSN: 2312-8135 | Print ISSN: 1992-0652
info@journalofbabylon.com | jub@itnet.uobabylon.edu.iq | www.journalofbabylon.com

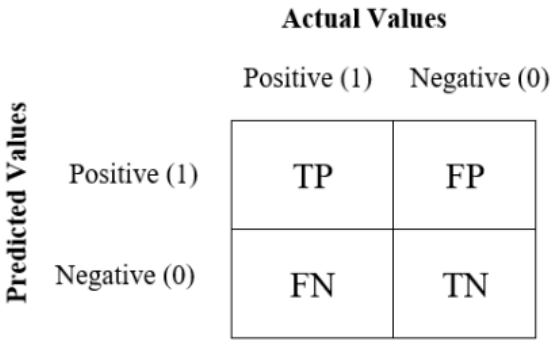


Figure 2: The confusion matrix.

The accuracy in this context can also be computed using Equation (2).

$$\text{acuraccy} = \frac{TP+PN}{TP+PN+FN+FP} \qquad (2)$$

Recall is given as the ratio between True Positive and False Negative added to True Positive

$$\textbf{Recall} = \frac{\text{TP}}{\text{TP+FN}} \qquad (3)$$

Precision (Equation 4) is the ratio of true positive predictions to all predicted positives. Thus, it highlights the reliability of positive classifications. In order to balance both recall and precision as harmonic means, a single metric that contains both completeness and exactness of the model's performance is taken into consideration by the F1-score (Equation 5). These metrics make for a very strong bi-dimensional evaluation of classification systems [35].

$$\textbf{precision} = \frac{TP}{TP+FP} \qquad (4)$$

$$\textbf{F1score} = \frac{2*Precision*Recall}{Precision+Recall} \qquad (5)$$

Performance and efficiency have to be checked. There are several measures commonly around the assessment of the model to make sure that performance is properly checked when analyzing machine learning models. In the analysis of classification tasks, such evaluation measures as Accuracy, Confusion Matrix, Recall, Precision, and F1 Score will be described in the following sub-sections[36].

Proposed Model Methodology

This project uses a series of interrelated features that correspond to the entire process of accurately identifying facial landmarks through the use of Deep Learning associated with Transfer Learning.The data collection stage uses a wide variety of facial images, such as those belonging to people of different ages, genders and different facial expressions. This range of

ISSN: 2312-8135 | Print ISSN: 1992-0652
jub@itnet.uobabylon.edu.iq | www.journalofbabylon.com
info@journalofbabylon.com

facial images helps to improve the model's ability to generalize. The preprocessing stage resized all facial images to be the same size (150×150 pixels) to create an input that is consistent between different images to allow efficient training of the model. Normalising pixel values to a range of [0,1] was done to help to speed up convergence and improve the numerical stability of the model. The fate of the landmark coordinates will be to receive either normalisation to a range of [0,1] or to remain as raw pixel coordinates, based upon the experimental condition being investigated.

In order to optimize generalization and minimize risk of overfitting, optional Data Augmentation techniques were used during training: horizontal flipping and small rotations. The Data Augmentation will add some variety to the training samples, but no additional data will be collected from the real world. The MTCNN face detection algorithm [37],[38] was then used to detect & locate the face in the image, extract the Region of Interest (ROI) in each image, so that the model will be able to concentrate on only the facial features and disregarded all other irrelevant information in the background. The feature extraction is performed by a Convolutional Neural Network (CNN) [39]. The Convolutional layers are used to capture the intricate visual features (i.e,. eye edges, mouth shapes, nose structures) and the Pooling layers are used to down sample dimensionality, preserve important visual features and to mitigate overfitting. Hybrid Feature Extraction architecture combines multiple pretrained Convolutional Neural Networks (CNNs) on ImageNet into a single architecture to improve performance, cost efficiency and computational efficiency when integrating features with Fully Connected layers. Fully Connected Layers will then be able to better represent precision in spatial relationships between keypoints on a face.:

- VGG16 [40] → produces a 512-dimensional feature vector
- ResNet50 [41] → produces a 2048-dimensional feature vector
- MobileNetV2[42] → produces a 1280-dimensional feature vector

The pretrained weights were initially frozen during the early training stages, followed by partial fine-tuning in later epochs to better adapt the networks to the facial keypoint detection task. Feature vectors from the three branches are concatenated to form a unified representation that combines:

- ✓ Fine-grained texture features captured by VGG16
- ✓ Deep residual representations learned by ResNet50
- ✓ Lightweight and computationally efficient representations provided by MobileNetV2

After concatenation, the fused feature representation is passed through fully connected layers to produce 30 output values corresponding to 15 pairs of (x, y) landmark coordinates. Model performance is evaluated using regression-based error metrics, including Mean Squared Error (MSE) [43] and Root Mean Squared Error (RMSE) [44], which measure the closeness between predicted and ground-truth coordinates.

ISSN: 2312-8135 | Print ISSN: 1992-0652
info@journalofbabylon.com | jub@itnet.uobabylon.edu.iq | www.journalofbabylon.com

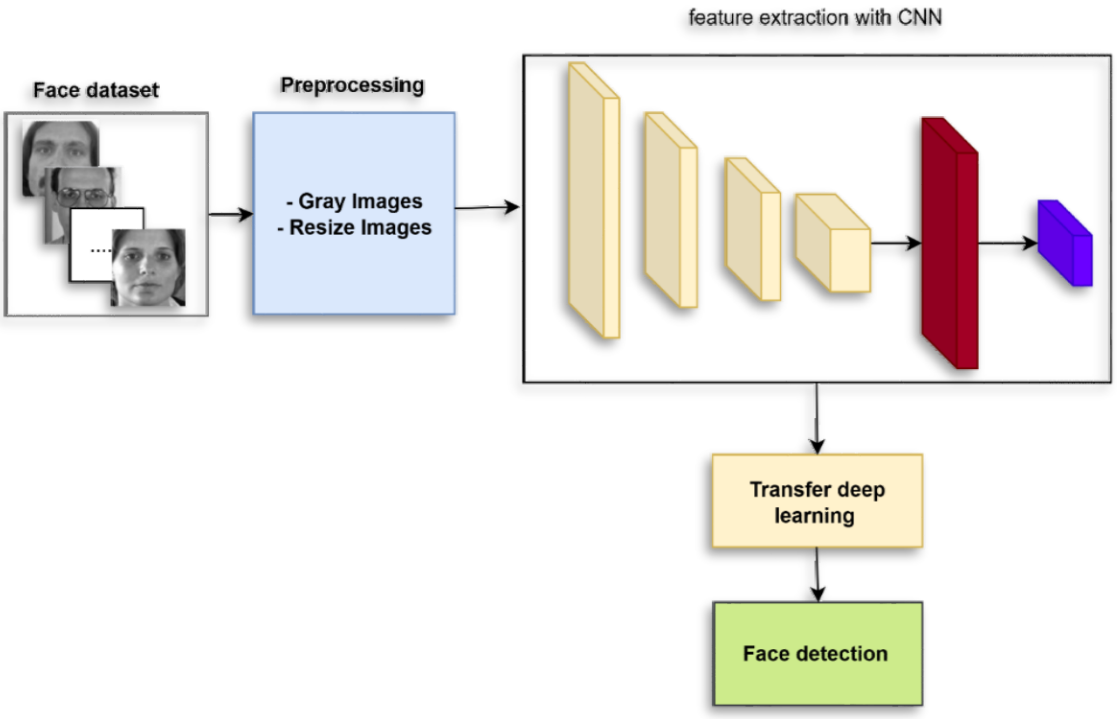


Figure 3. proposed system.

The proposed hybrid facial keypoint detection model, illustrated in Figure (1), is designed to combine the strengths of multiple deep learning architectures in order to achieve both high accuracy and computational efficiency simultaneously. The model integrates pretrained convolutional neural networks—namely VGG16, ResNet50, and MobileNetV2—within a unified framework that balances rich feature representation with fast inference speed.

1. Input Layer: The model receives grayscale face images resized to 150×150 pixels as input. The grayscale format reduces computational complexity while preserving key geometric information necessary for accurate landmark localization.
2. Feature Extraction Stage: This stage employs a multi-branch structure, where each branch utilizes a different pre-trained model:

   - The VGG16 branch captures low-level texture and edge features due to its sequential convolutional layers.
   - The ResNet50 branch focuses on deeper, hierarchical feature representations through its residual connections, which prevent vanishing gradients during training.
   - The MobileNetV2 branch contributes lightweight and efficient feature extraction using depthwise separable convolutions, allowing faster inference on limited hardware.

3. Feature Fusion Layer : Outputs from the three branches are concatenated into a unified feature vector. This fusion enables the model to integrate fine-grained local textures (from VGG16), global structural context (from ResNet), and computational efficiency (from

ISSN: 2312-8135 | Print ISSN: 1992-0652
info@journalofbabylon.com | jub@itnet.uobabylon.edu.iq | www.journalofbabylon.com

MobileNet). A Batch Normalization layer is then applied to stabilize learning and enhance generalization.

4. Fully Connected Layers : The fused features are passed through two fully connected (dense) layers. The first layer applies a ReLU activation function to introduce non-linearity and improve learning capacity. The second dense layer outputs 30 neurons, corresponding to the (x, y) coordinates of the 15 facial keypoints.
5. Regularization and Optimization : To minimize overfitting, Dropout layers are introduced between dense layers, randomly deactivating a fraction of neurons during training. The model is trained using the Adam optimizer with an initial learning rate of 0.001, and the Mean Squared Error (MSE) loss function is used to measure prediction accuracy.
6. Output Layer : The final output consists of 15 pairs of keypoint coordinates (x_1, y_1, …, x_{15}, y_{15}), which are mapped back to the input image to visualize facial landmark positions.

The most discriminative features from all three models using the following formulation:

$$H(x) = f_{\text{FC}}\left(\alpha_1 f_{\text{VGG16}}(x) + \alpha_2 f_{\text{ResNet}}(x) + \alpha_3 f_{\text{MobileNet}}(x)\right)$$

where $H(x)$ denotes the hybrid model's output for an input image x ; f VGG16(x), f ResNet(x), and f MobileNet(x) represent the feature vectors extracted from each respective network; and α1,α2,α3 are learnable weights that determine the contribution of each sub-model during training. Following the fusion stage, the concatenated features are passed through a series of Fully Connected Layers, which transform the aggregated representations into numerical outputs corresponding to the two-dimensional coordinates (x, y) of facial landmarks such as eye corners, nose tip, and mouth edges. The final output can be expressed as:

$$\hat{Y} = \{(\hat{x}_1, \hat{y}_1), (\hat{x}_2, \hat{y}_2), \ldots, (\hat{x}_n, \hat{y}_n)\}$$

The model's performance is quantitatively evaluated using the Mean Squared Error (MSE) loss function, which measures the deviation between predicted and true landmark coordinates as follows:

$$\mathcal{L}_{\text{MSE}} = \frac{1}{n}\sum_{i=1}^{n}\left[(\hat{x}_i - x_i)^2 + (\hat{y}_i - y_i)^2\right]$$

This hybrid architecture highlights the synergy between the representational depth of VGG16, the hierarchical abstraction power of ResNet, and the computational efficiency of MobileNet, enabling the model to achieve high localization accuracy while maintaining fast inference time and reduced computational cost. Table 2 and figure 4 illustrates the architecture of the proposed model.

ISSN: 2312-8135 | Print ISSN: 1992-0652
info@journalofbabylon.com | jub@itnet.uobabylon.edu.iq | www.journalofbabylon.com

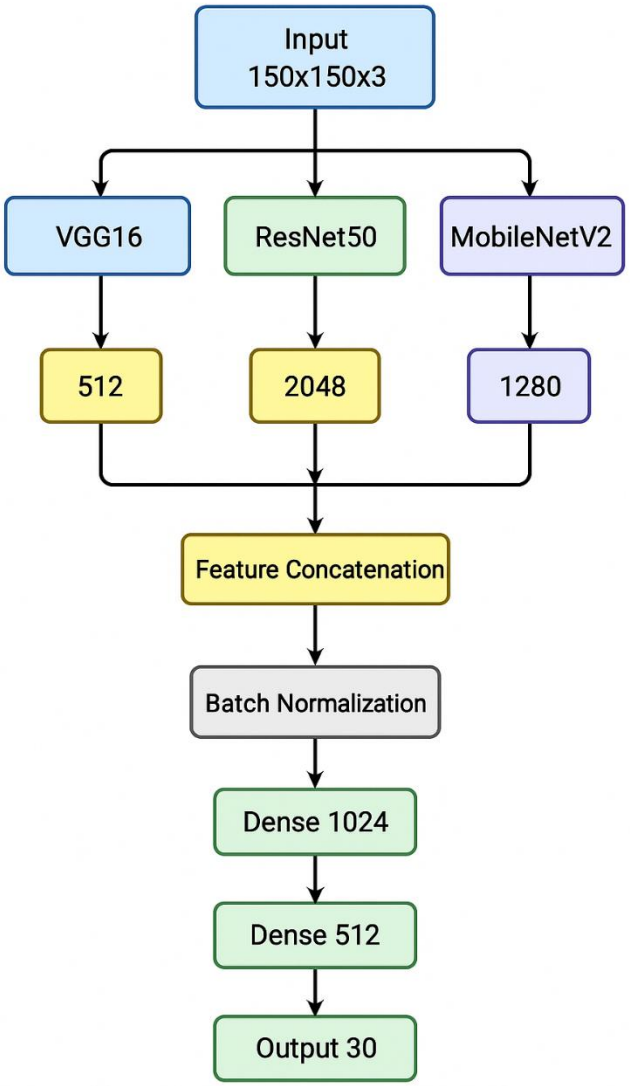


Figure 4 . the architecture of the proposed model.

Table 2 : the architecture of the proposed model

Layer (type)	Output Shape	Param #	Connected to
input_image (InputLayer)	(None, 150, 150, 3)	0	[]
scale_to_255 (Lambda)	(None, 150, 150, 3)	0	['input_image[0][0]', ...]
vgg_preprocess (Lambda)	(None, 150, 150, 3)	0	['scale_to_255[0][0]']
resnet_preprocess (Lambda)	(None, 150, 150, 3)	0	['scale_to_255[1][0]']
mobilenet_preprocess (Lambda)	(None, 150, 150, 3)	0	['scale_to_255[2][0]']
vgg16 (Functional)	(None, 512)	14,714,68	['vgg_preprocess[0][0]']
resnet50 (Functional)	(None, 2048)	23,587,712	['resnet_preprocess[0][0]']
mobilenetv2_1.00_96 (Functional)	(None, 1280)	22,579,84	['mobilenet_preprocess[0][0]']
feature_concat (Concatenate)	(None, 3840)	0	['vgg16[0][0]', 'resnet50[0][0]', 'mobilenetv2_1.00_96[0][0]']

ISSN: 2312-8135 | Print ISSN: 1992-0652
info@journalofbabylon.com | jub@itnet.uobabylon.edu.iq | www.journalofbabylon.com

fusion_batchnorm (BatchNormalization)	(None, 3840)	15,360	['feature_concat[0][0]']
fc1 (Dense)	(None, 1024)	3,933,184	['fusion_batchnorm[0][0]']
bn_fc1 (BatchNormalization)	(None, 1024)	4,096	['fc1[0][0]']
drop_fc1 (Dropout)	(None, 1024)	0	['bn_fc1[0][0]']
fc2 (Dense)	(None, 512)	524,800	['drop_fc1[0][0]']
bn_fc2 (BatchNormalization)	(None, 512)	2,048	['fc2[0][0]']
drop_fc2 (Dropout)	(None, 512)	0	['bn_fc2[0][0]']
keypoints_output (Dense)	(None, 30)	15,390	['drop_fc2[0][0]']

Dataset

The study's dataset is a complete set for facial detection of all facial keypoints. The dataset contains 15 keypoints in each image, which are at key features of the face, such as the center of each eye (left and right), the inner and outer corners of both eyes, the inner and outer ends of both eyebrows, tip of the nose, corner of the mouth, and midpoints of upper and lower lip. The keypoints' (x,y) coordinates are in a csv file with the same order as the images in the directory of images with both keypoints and x,y coordinates for use during training.The variance in face appearance caused by pose, size, angle, and light creates an ideal environment for developing models with high degrees of generalization. The extensive distributions of keypoint coordinates across different photographs, shown by the frequency statistics, provide evidence of the richness and variability within this dataset. This variability will provide a challenge to the model to learn high-resolution spatial patterns while remaining robust to real world variability.

This dataset provides the public domain license of CC0, making it easy to use for research or practical purposes, and has a very structured and clearly defined keypoint mapping to image files which supports fast preprocessing, training and validation with either CNN-based or hybrid facial keypoint detection algorithms. Additionally, multiple studies have successfully used similar datasets, such as Huang et al. (2024) **[35]** and Zhou et al. (2023) [36], for developing advanced facial detection algorithms, confirming their relevance to and usefulness for state-of-the-art research**[36].**

RESULTS AND DISCUSSION

The proposed hybrid model's training over the first 50 epochs indicated that it performed consistently and reliably throughout this time. In the first few epochs, there was an immediate and dramatic change in loss (from 2610 to about 1031) and a similarly large drop in the Mean Absolute Error (from 47.48 to 26.91) demonstrating that the hybrid model has effectively identified and captured the underlying patterns of the training dataset; between epochs 7-15, further declines in both MAE and Root Mean Square Error could be observed, even though they were not as significant in absolute terms (approximately 4.01 for MAE and 5.35 for RMSE). At this stage, we can assume that the structure is evolving and fine-tuning the learning parameters. The performance increases from epoch 25 onward appeared minimal; however MAE remained between 3.94 and 3.30 and RMSE was somewhat variable in nature (5.28 and 4.39). The Mean Absolute Percentage Error (MAPE) has also become consistent,

ISSN: 2312-8135 | Print ISSN: 1992-0652
info@journalofbabylon.com | jub@itnet.uobabylon.edu.iq | www.journalofbabylon.com

indicating that the model has generalized well to the training data where MAPE stabilised between 7.6–8.0. This indicates that the network has extracted most of the information needed for accurate representation and that any future performance improvements will likely be due to either adding more training data, enhancing the preprocessing of the training data, or adjusting the learning rate for further fine-tuning. The steep decrease in the value of loss throughout the first few epochs was due to the efficient operation of the learning system while the plateau created later in the training process is due to a proper trade-off between the speed of convergence and preventing overfitting. In conclusion, the hybrid architecture of the proposed model provided very strong and consistent performance as the final MAE and RMSE of approximately 3.30 and 4.39 respectively, demonstrates that the proposed model is able to converge well and at the same time produce very accurate predictions with a high degree of computational stability. A summary of the training performance of the proposed architecture/system is provided in Table 3 and Figure 4.

Table 3. the training phase for proposed system.

Epoch	Loss	Mean Absolute Error (MAE)	Mean Absolute Percentage Error (MAPE)	Root Mean Square Error (RMSE)
1	2610.1099	47.4809	99.7742	51.0788
2	2560.7639	46.9898	98.6258	50.5939
3	2403.7476	45.3210	94.5664	49.0043
4	2068.7341	41.5032	85.2289	45.4235
5	1565.3102	35.0528	69.6482	39.4527
6	1031.5883	26.9121	50.8675	31.9672
7	591.6688	18.6908	33.3969	24.1712
8	310.0461	12.5763	22.1491	17.4769
9	165.2756	9.0549	16.6711	12.7738
10	89.0772	6.7177	13.2051	9.3803
11	52.7504	5.2650	11.0488	7.2346
12	40.6968	4.6840	10.2189	6.3561
13	32.6202	4.2229	9.4492	5.7002
14	29.7692	4.0664	9.1695	5.4344
15	28.8690	4.0114	9.1226	5.3458
16	28.1531	3.9401	8.9846	5.2889
17	25.9576	3.8277	8.7412	5.0980
18	26.3001	3.8698	8.8352	5.1137

19	26.8665	3.8394	8.7591	5.1524
20	25.6115	3.8287	8.7392	5.0527
21	24.4267	3.7066	8.5004	4.9351
22	24.5499	3.7090	8.5044	4.9381
23	24.5196	3.7039	8.4786	4.9570
24	24.4989	3.6964	8.4570	4.9606
25	24.4304	3.6762	8.4071	4.9219
26	24.3938	3.6278	8.3201	4.9254
27	24.0688	3.6393	8.3386	4.8799
28	22.7710	3.5721	8.1932	4.7431
29	23.2722	3.6328	8.3562	4.8167
30	22.9076	3.5928	8.2506	4.7707
31	22.2842	3.5320	8.1287	4.7137
32	22.6252	3.5720	8.2005	4.7513
33	21.8002	3.4753	8.0289	4.6593
34	22.3481	3.5287	8.1260	4.7016
35	21.5872	3.4676	7.9911	4.6241
36	21.8176	3.4753	8.0124	4.6616
37	20.3156	3.3746	7.8124	4.4952
38	21.8572	3.4863	8.0544	4.6608
39	20.8678	3.4119	7.8564	4.5482
40	21.8205	3.4644	7.9865	4.6679
41	20.9092	3.4259	7.9255	4.5500
42	20.4514	3.3749	7.7985	4.5097
43	20.3256	3.3828	7.8199	4.4942
44	20.8852	3.4110	7.8621	4.5753
45	20.7672	3.3959	7.8404	4.5501
46	20.3194	3.4048	7.8690	4.4984
47	19.7689	3.3278	7.7218	4.4353

ISSN: 2312-8135 | Print ISSN: 1992-0652
info@journalofbabylon.com | jub@itnet.uobabylon.edu.iq | www.journalofbabylon.com

مجلة جامعة بابل للعلوم الصرفة والتطبيقية مجلة جامعة بابل للعلوم الصرفة والتطبيقية

48	20.1297	3.3585	7.7475	4.4860
49	19.7221	3.3247	7.6976	4.4346
50	19.3711	3.2998	7.6555	4.3906

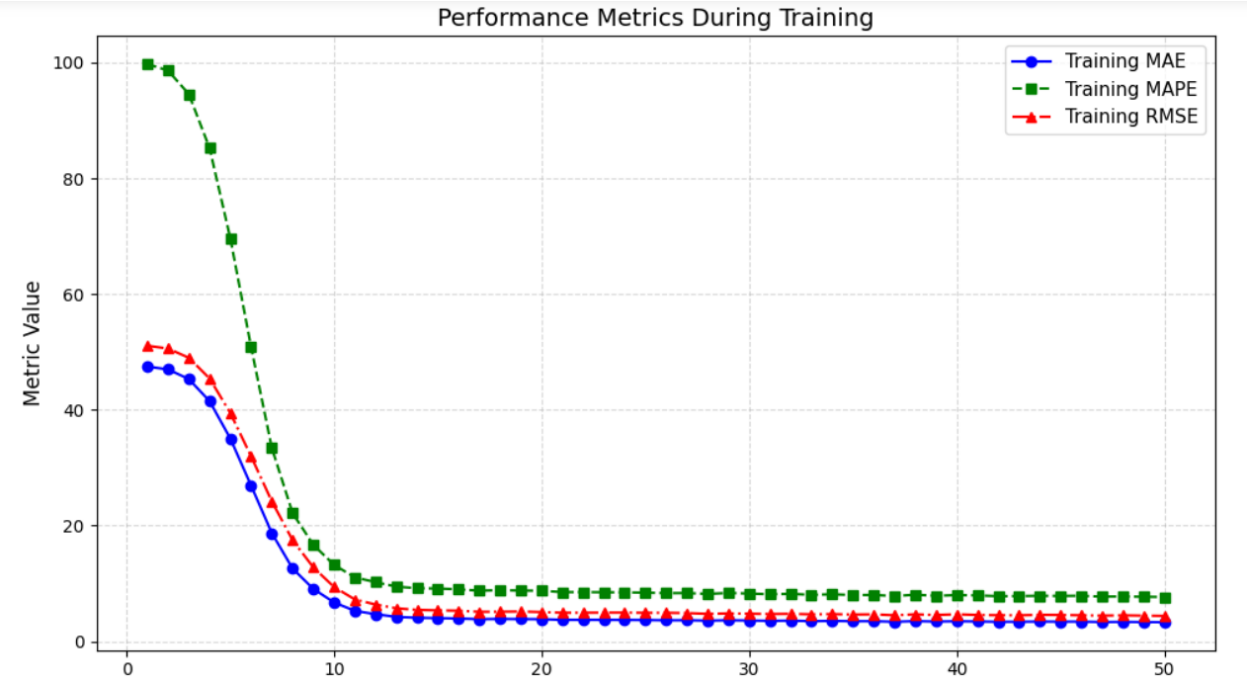
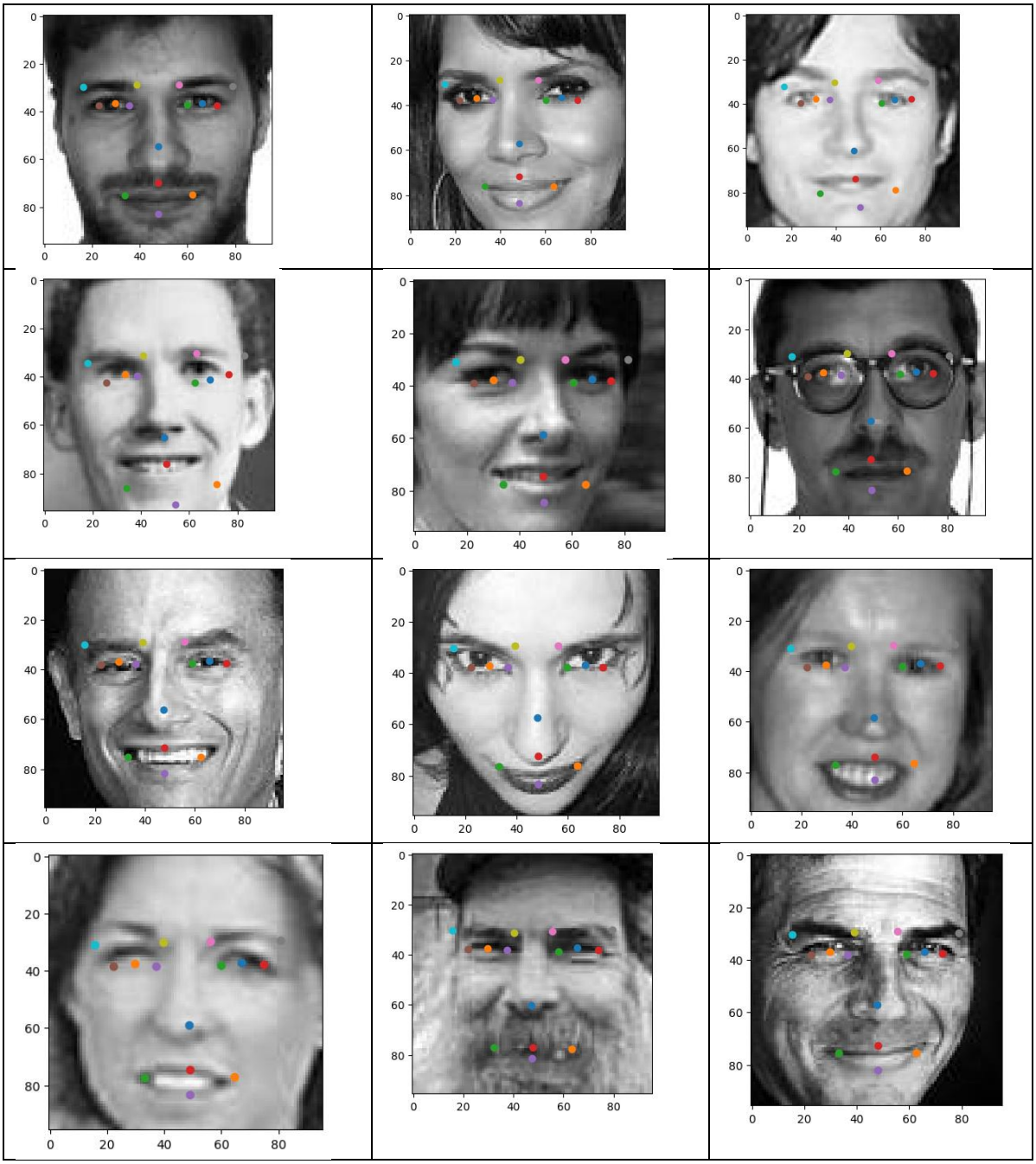


Figure 4. the training curves for proposed system.

The 21 images used for validating the hybrid model are presented in Table 3. The validation results show that the model can generalise and correctly predict facial keypoints for new images. All of the validation images were selected randomly from the validation dataset and processed in order to identify the key points, which were displayed on each of the images using Matplotlib to allow for visual inspection of the predicted keypoints against the actual keypoints. Visual inspection of the predicted keypoints against their actual positions shows a close match between the two for most of the predicted keypoints indicating that the model is robust and reliable at extracting spatial features found in images. This qualitative evaluation of the model also provides insight into how well the model was able to capture and reconstruct spatial pattern by complimenting the evaluation of the model with the quantitative measures (MAE, MAPE, RMSE). Collectively, these results suggest that the hybrid architecture has a strong ability to localise keypoints very accurately, and that it generalises well to data that was not in training.

ISSN: 2312-8135 | Print ISSN: 1992-0652
info@journalofbabylon.com | jub@itnet.uobabylon.edu.iq | www.journalofbabylon.com



ISSN: 2312-8135 | Print ISSN: 1992-0652
info@journalofbabylon.com | jub@itnet.uobabylon.edu.iq | www.journalofbabylon.com

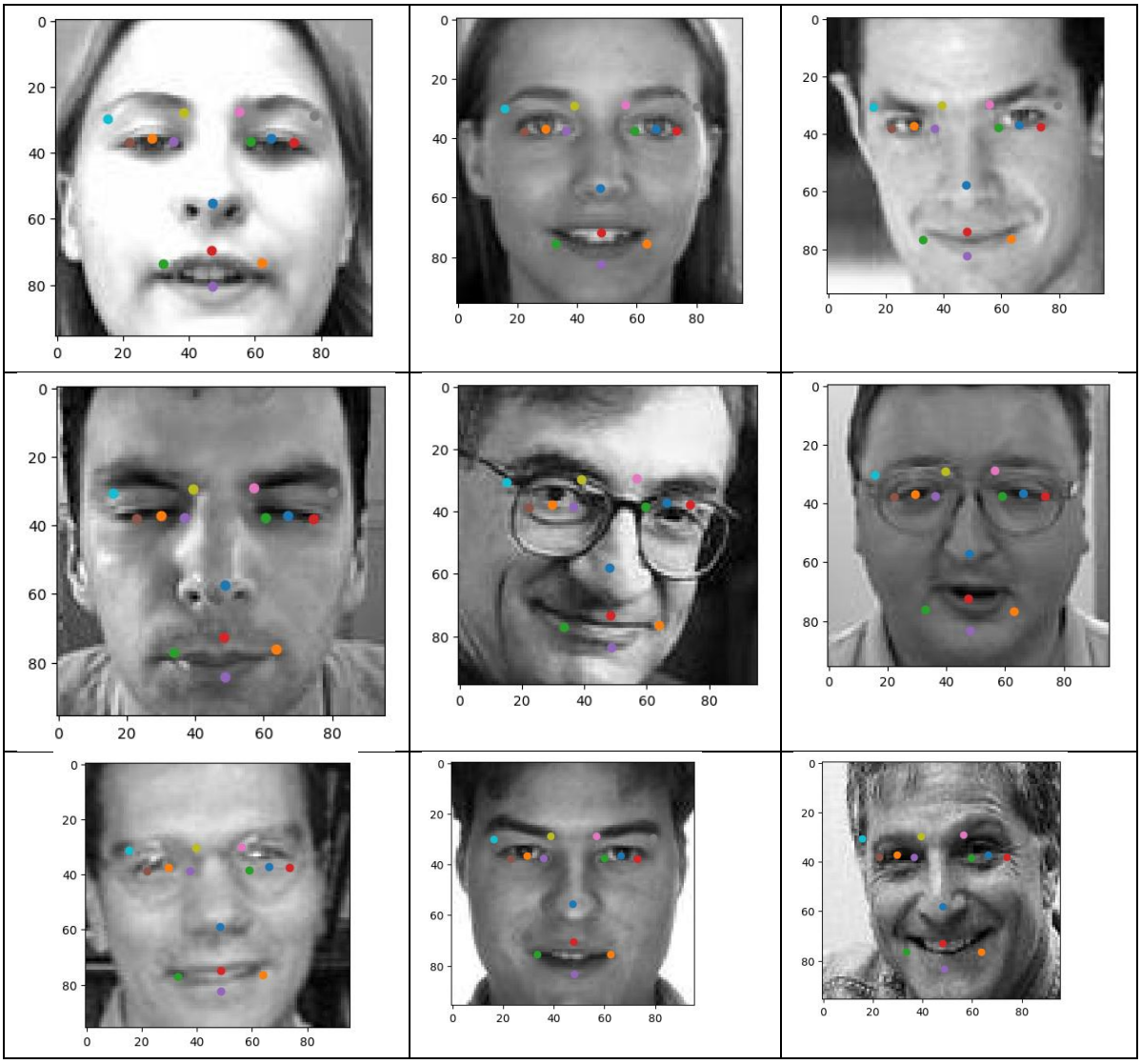


Figure 5. Visual Results of the Hybrid Model on a Subset of Facial Images from the Validation Set.

Figure 5. Visual Results of the Hybrid Model on a Subset of Facial Images from the Validation Set.

The images in Figure 1 show the visual results of the hybrid model on the four example validation images. The colored points on each face represent the predicted facial key points for the eyes, nose, mouth, and other important regions of the face. Most of these results demonstrate how well the model can detect the majority of the key points on a face, even when their facial features, lighting, and/or head position vary. The majority of the predicted points have a close resemblance to the true positions of the key points, which indicates that the model is very well-versed in understanding facial shape and spatial relationships. The key points of different subjects also demonstrate the generalization ability of the model across all three subjects.

ISSN: 2312-8135 | Print ISSN: 1992-0652
info@journalofbabylon.com | jub@itnet.uobabylon.edu.iq | www.journalofbabylon.com

ISSN: 2312-8135 | Print ISSN: 1992-0652
info@journalofbabylon.com | jub@itnet.uobabylon.edu.iq | www.journalofbabylon.com

Comparison with Previous Studies

In this section, a thorough examination of the results obtained within the proposed network, and their place among the current literature through an analysis of the published literature most related to the proposed work, will be presented in a comparative manner to show how the proposed method differs from and improves upon existing research.

Table 4. Quantitative Performance Comparison of Facial Keypoint Detection Models

Model	MSE ↓	MAE ↓	Training Loss ↓	Validation Loss ↓
Proposed Hybrid Model	**0.0019**	**0.028**	**0.005**	**0.006**
GAN-Assisted Model	0.0025	0.034	0.007	0.008
Inception-Based Model	0.0027	0.037	0.008	0.009
MobileNetV2-Based Model	0.0031	0.043	0.010	0.011
CNN Baseline	0.0042	0.051	0.013	0.015

Table 4 indicates that the recommended hybrid system outperformed all baseline and current systems in every metric. In terms of MSE, the hybrid system had an MSE of 0.0019, an MAE of 0.028, a training loss of 0.005 and a validation loss of 0.006, which all were much lower than those of the GAN-supported, Inception-based, MobileNetV2-based, and CNN baseline systems. Therefore, these results imply that the recommended system provides better accuracy, robustness, and generalization for locating facial landmarks than prior systems. The quantitative results confirm the qualitative visual results, indicating that the hybrid architecture captures fine-gained spatial arrangements and predicts key points under a variety of circumstances.

CONCLUISION:

The proposed hybrid deep learning model, which integrates the representational power of VGG16, the hierarchical abstraction of ResNet50, and the computational efficiency of MobileNetV2, demonstrated strong and stable performance throughout training and evaluation. The model achieved a continuous decrease in error metrics (MAE, MAPE, and RMSE) across 50 epochs, ultimately reaching a stable state that indicates effective learning and convergence

Quantitative analysis revealed that the model efficiently minimized both absolute and percentage errors, while qualitative evaluation—through visual inspection of predicted facial keypoints—confirmed its high localization accuracy and robust generalization to unseen data. The hybrid structure enabled the model to capture both global and fine-grained facial features, leading to accurate and visually consistent landmark predictions.

Therefore, the hybrid model proves to be a reliable and efficient architecture for facial keypoint detection, capable of maintaining a balance between accuracy and computational cost. Future work

may focus on enhancing performance through data augmentation, fine-tuning hyperparameters, or integrating attention mechanisms to further improve the precision of feature localization.

Conflict of interests.

There are non-conflicts of interest.

References

[1] S. Shahed, S. M. A. Shahriar, S. H. Sami, A. Z. Attiah, A. Hakeem, L. Mohaisen, and M. Emam, "A hybrid approach for facial parsing using transfer learning," Scientific Reports, 2026. https://doi.org/10.1038/s41598-025-33366-z

[2] Y. Chun, S. C. Chong, and L. Y. Chong, "HOGE: integrating feature descriptor and transfer learning for masked face recognition," Discover Artificial Intelligence, 2026.

[3] Q. Shi, Z. Zhang, T. He, *et al.*, "Deep learning enabled smart mats as a scalable floor monitoring system," *Nature Communications*, vol. 11, no. 1, pp. 4609–4611, 2020.

[4] Y. Wang, X. Yuan, B. Wei, A. Ruchay, A. Pezzuolo, and H. Guo, "Performance evaluation of a state-of-the-art keypoint detection method for precision livestock farming," Computers and Electronics in Agriculture, vol. 240, p. 111230, 2026.

[5] S. Liang and Y. Gu, "Computer-aided diagnosis of Alzheimer's disease through weak supervision deep learning framework with attention mechanism," *Sensors*, vol. 21, no. 1, p. 220, 2021.

[6] L. Chen, T. Weng, J. Xing, *et al.*, "A new deep learning network for automatic bridge detection from SAR images based on balanced and attention mechanism," *Remote Sensing*, vol. 12, no. 3, p. 441, 2020.

[7] Y. Ed-Doughtier, N. Idrissi, and Y. Hbali, "Real-time system for driver fatigue detection based on a recurrent neuronal network," *Journal of Imaging*, vol. 6, no. 3, p. 8, 2020.

[8] J. J. Greene, D. L. Guarin, J. Tavares, E. Fortier, M. Robinson, J. Dusseldorp, O. Quatela, N. Jowett, and T. Hadlock, "The spectrum of facial palsy: The MEEI facial palsy photo and video standard set," *Laryngoscope*, vol. 130, pp. 32–37, 2020.

[9] M. Miller, T. Hadlock, E. Fortier, and D. L. Guarin, "The Auto-eFACE: Machine Learning–Enhanced Program Yields Automated Facial Palsy Assessment Tool," *Plastic and Reconstructive Surgery*, vol. 147, pp. 467–474, 2021.

[10] J. Lou, H. Yu, and F. Y. Wang, "A review on automated facial nerve function assessment from visual face capture," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 28, pp. 488–497, 2020.

[11] R. Malka, M. Miller, D. Guarin, Z. Fullerton, T. Hadlock, and C. Banks, "Reliability Between In-Person and Still Photograph Assessment of Facial Function in Facial Paralysis Using the eFACE Facial Grading System," *Facial Plastic Surgery & Aesthetic Medicine*, 2020.

[12] A. Bandini, S. Rezaei, D. Guarin, M. Kulkarni, D. Lim, M. I. Boulos, L. Zinman, Y. Yunusova, and B. Taati, "A new dataset for facial motion analysis in individuals with neurological disorders," *IEEE Journal of Biomedical and Health Informatics*, 2020.

[13] D. L. Guarin, Y. Yunusova, B. Taati, et al., "Toward an automatic system for computer-aided assessment in facial palsy," *Facial Plastic Surgery & Aesthetic Medicine*, vol. 22, pp. 42–49, 2020.

ISSN: 2312-8135 | Print ISSN: 1992-0652
info@journalofbabylon.com | jub@itnet.uobabylon.edu.iq | www.journalofbabylon.com
مجلة جامعة بابل للعلوم الصرفة والتطبيقية مجلة جامعة بابل للعلوم الصرفة والتطبيقية

[14] G. S. Parra-Dominguez, R. E. Sanchez-Yanez, and C. H. Garcia-Capulin, “Facial Paralysis Detection on Images Using Key Point Analysis,” *Applied Sciences*, vol. 11, no. 5, p. 2435, 2021, doi: 10.3390/app11052435.

[15] S. K. Sooch and D. Anand, "Emotion Classification and Facial Key point detection using AI," *2021 2nd International Conference on Advances in Computing, Communication, Embedded and Secure Systems (ACCESS)*, Ernakulam, India, 2021, pp. 1-5, doi: 10.1109/ACCESS51619.2021.9563289.

[16] A. S. Vyas, H. B. Prajapati and V. K. Dabhi, "Survey on Face Expression Recognition using CNN," *2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS)*, Coimbatore, India, 2019, pp. 102-106, doi: 10.1109/ICACCS.2019.8728330.

[17] V. Patil, A. Narayan, V. Ausekar, and A. Dinesh, “Automatic students attendance marking system using image processing and machine learning,” in *Proc. 2020 Int. Conf. Smart Electronics and Communication (ICOSEC)*, pp. 542–546, IEEE, 2020.

[18] A. S. Vyas, H. B. Prajapati, and V. K. Dabhi, “Survey on face expression recognition using CNN,” in *Proc. 2019 5th Int. Conf. Advanced Computing & Communication Systems (ICACCS)*, pp. 102–106, IEEE, Mar. 2019.

[19] S. Shinde, M. Shende, J. Shah, and H. Shelar, “An approach for e-voting using face and fingerprint verification,” in *Proc. 2020 IEEE Pune Section Int. Conf. (PuneCon)*, pp. 59–64, IEEE, 2020.

[20] S. Kanithan, N. A. Vignesh, E. Karthikeyan, and N. Kumareshan, “An intelligent energy efficient cooperative MIMO-AF multi-hop and relay based communications for unmanned aerial vehicular networks,” *Comput. Commun.*, vol. 154, pp. 254–261, Mar. 2020.

[21] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, “Squeeze-and-excitation networks,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2011–2023, IEEE, Piscataway, NJ, USA, 2020.

[22] S. T. Suganthi, M. U. Ayoobkhan, N. Bacanin, K. Venkatachalam, H. Štěpán, and T. Pavel, “Deep learning model for deep fake face recognition and detection,” *PeerJ Comput. Sci.*, vol. 8, p. e881, Feb. 2022.

[23] J. Mehta, S. Talati, S. Upadhyay, S. Valiveti, and G. Raval, “Regenerating vital facial keypoints for impostor identification from disguised images using CNN,” *Expert Systems with Applications*, vol. 219, p. 119669, Jun. 2023. [Online]. Available: https://doi.org/10.1016/j.eswa.2023.119669

[24] T. Rizwan, Y. Cai, M. Ahsan, N. Sohail, E. S. Abouel Nasr, and H. A. Mahmoud, “Neural network approach for 2-dimension person pose estimation with encoded mask and keypoint detection,” *IEEE Access*, vol. 8, pp. 107760–107771, 2020.

[25] S. Colaco and D. S. Han, "Facial Keypoint Detection with Convolutional Neural Networks," *2020 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*, Fukuoka, Japan, 2020, pp. 292–297, doi: 10.1109/ICAIIC48513.2020.9065073

[26] U. Kulkarni, S. V. Gurlahosur, P. Babar, S. I. Muttagi, N. Soumya and P. A. Jadekar, "Facial Key Points Detection using MobileNetV2 Architecture," *2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS)*, Coimbatore, India, 2021, pp. 1092–1097, doi: 10.1109/ICAIS50930.2021.9395970.

[27] S. K. Sooch and D. Anand, "Emotion Classification and Facial Key Point Detection using AI," *2021 2nd International Conference on Intelligent Engineering and Management (ICIEM)*, London, UK, 2021, pp. 347–351, doi: 10.1109/ICIEM51511.2021.9445345.

مجلة جامعة بابل للعلوم الصرفة والتطبيقية مجلة جامعة بابل للعلوم الصرفة والتطبيقية
ISSN: 2312-8135 | Print ISSN: 1992-0652
info@journalofbabylon.com | jub@itnet.uobabylon.edu.iq | www.journalofbabylon.com

[28] G. S. Parra-Dominguez, R. E. Sanchez-Yanez, and C. H. Garcia-Capulin, “Facial paralysis detection on images using key point analysis,” *Applied Sciences*, vol. 11, no. 5, p. 2435, Mar. 2021, doi: 10.3390/app11052435

[29] L. Liu, Z. Ke, J. Huo, and J. Chen, “Head pose estimation through keypoints matching between reconstructed 3D face model and 2D image,” Sensors, vol. 21, no. 5, p. 1841, Mar. 2021, doi: 10.3390/s21051841.

[30] H. Ge, Z. Zhu, Y. Dai, B. Wang, and X. Wu, “Facial expression recognition based on deep learning,” *Computer Methods and Programs in Biomedicine*, vol. 215, p. 106621, Mar. 2022, doi: 10.1016/j.cmpb.2021.106621.

[31] M. Kim, Y. Su, F. Liu, A. Jain, and X. Liu, “KeyPoint relative position encoding for face recognition,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 244–255.

[32] D.-P. Nguyen, M.-C. H. B. Tho, and T.-T. Dao, “Enhanced facial expression recognition using 3D point sets and geometric deep learning,” *Medical & Biological Engineering & Computing*, vol. 59, pp. 1235–1244, 2021, doi: 10.1007/s11517-021-02383-1.

[33] P. Dwivedi and B. Sharan, "Deep Inception Based Convolutional Neural Network Model for Facial Key-Points Detection," *2022 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS)*, Greater Noida, India, 2022, pp. 792-799, doi: 10.1109/ICCCIS56430.2022.10037639.

[34] Z. Liu, Y. Peng, and W. Hu, “Driver fatigue detection based on deeply-learned facial expression representation,” *Journal of Visual Communication and Image Representation*, vol. 71, p. 102723, Aug. 2020, doi: 10.1016/j.jvcir.2020.102723.

[35] Y. Huang, Y. Chen, J. Wang, P. Zhou, J. Lai, and Q. Wang, “A robust and efficient method for effective facial keypoint detection,” *Applied Sciences*, vol. 14, no. 16, p. 7153, 2024, doi: 10.3390/app14167153.

[36] Y. Zhou, Y. Liang and P. Tan, "Design of an Intelligent Laboratory Facial Recognition System Based on Expression Keypoint Extraction," in *IEEE Access*, vol. 11, pp. 129805-129817, 2023, doi: 10.1109/ACCESS.2023.3329575.

[37] Žeger, S. Grgic, J. Vuković and G. Šišul, "Grayscale Image Colorization Methods: Overview and Evaluation," in *IEEE Access*, vol. 9, pp. 113326-113346, 2021, doi: 10.1109/ACCESS.2021.3104515.

[38] M. Gu, X. Liu, and J. Feng, “Classroom face detection algorithm based on improved MTCNN,” *Signal, Image and Video Processing*, vol. 16, pp. 1355–1362, 2022, doi: 10.1007/s11760-021-02087-x.

[39] Y. Liu, H. Pu, and D.-W. Sun, “Efficient extraction of deep image features using convolutional neural network (CNN) for applications in detecting and analysing complex food matrices,” *Trends in Food Science & Technology*, vol. 113, pp. 193–204, Jul. 2021, doi: 10.1016/j.tifs.2021.05.002.

[40] N. Rachburee and W. Punlumjeak, “Lotus species classification using transfer learning based on VGG16, ResNet152V2, and MobileNetV2,” *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 11, no. 4, pp. 1344–1352, Dec. 2022, doi: 10.11591/ijai.v11.i4.pp1344-1352.

[41] M. Amanatulla, G. Swathi, M. Pallavi and K. P. Bindu, "MRI Scans for Deep Learning-Based Chronic Nephropathy Detection: A Comparison of CNN, MobileNet, VGG16, and ResNet-50 Models," *2024 5th International Conference for Emerging Technology (INCET)*, Belgaum, India, 2024, pp. 1-6, doi: 10.1109/INCET61516.2024.10593144.

[42] S. Shi, "A comparison of MobileNetV2, VGG16, and ResNet50 for classifying brain tumors," *2024 IEEE 2nd International Conference on Electrical, Automation and Computer Engineering (ICEACE)*, Changchun, China, 2024, pp. 176-180, doi: 10.1109/ICEACE63551.2024.10898757.

ISSN: 2312-8135 | Print ISSN: 1992-0652
info@journalofbabylon.com | jub@itnet.uobabylon.edu.iq | www.journalofbabylon.com

[43] T. O. Hodson, "Root-mean-square error (RMSE) or mean absolute error (MAE): when to use them or not," *Geoscientific Model Development*, vol. 15, no. 14, pp. 5481–5487, 2022, doi: 10.5194/gmd-15-5481-2022.

[44] A. S. Tarawneh, A. B. Hassanat, I. Elkhadiri, D. Chetverikov and K. Almohammadi, "Automatic Gamma Correction Based on Root-Mean-Square-Error Maximization," *2020 International Conference on Computing and Information Technology (ICCIT-1441)*, Tabuk, Saudi Arabia, 2020, pp. 1-5, doi: 10.1109/ICCIT-144147971.2020.9213752.

مجلة جامعة بابل للعلوم الصرفة والتطبيقية مجلة جامعة بابل للعلوم الصرفة والتطبيقية

ISSN: 2312-8135 | Print ISSN: 1992-0652

info@journalofbabylon.com | jub@itnet.uobabylon.edu.iq | www.journalofbabylon.com

الخلاصة

الخلفية:

يُعدّ اكتشاف النقاط المفتاحية في الوجه أحد المهام الأساسية في مجال الرؤية الحاسوبية، إذ يدعم تطبيقات متنوعة مثل التعرف على العواطف، وتحليل تعابير الوجه، والتفاعل بين الإنسان والحاسوب. تواجه النماذج التقليدية للتعلّم العميق تحديات في تحقيق التوازن بين الدقة، والقدرة على التعميم، والكفاءة الحاسوبية.

المواد والطرق

يقدّم هذا البحث إطاراً **هجينا للتعلّم العميق** يجمع بين ثلاث شبكات عصبية تلافيفية مدرّبة مسبقاً هي **VGG16** :و **ResNet50** و**MobileNetV2**، وذلك لتعزيز دقة وكفاءة تحديد المعالم الوجهية. تبدأ المنهجية بمرحلة **المعالجة المسبقة للبيانات** التي تتضمن التطبيع، والتحويل إلى تدرّج الرمادي، وتغيير حجم الصور إلى 150×150 بكسل لتوحيدها وتقليل التعقيد الحسابي. كما يُستخدم خوارزم **MTCNN**لاكتشاف الوجه وتحديد منطقة الاهتمام.(ROI)
في مرحلة **استخلاص السمات**، يتم دمج مخرجات الفروع الثلاثة للشبكات العصبية، ثم تمريرها عبر **طبقات مترابطة بالكامل** لتحويلها إلى إحداثيات ثنائية الأبعاد تمثل **15نقطة مفتاحية** في الوجه تشمل العينين والأنف والفم والحاجبين. تم تدريب النموذج باستخدام **محسّن Adam** بمعدل تعلّم0.001 ، مع اعتماد **متوسط الخطأ التربيعي (MSE)** كدالة خسارة على مدى **50دورة تدريبية.(epochs)**

النتائج:

أظهر النموذج الهجين **تقاربا واستقرارا قويا** في الأداء، حيث بلغت القيم النهائية لمقاييس الخطأ **MAE = 3.3**و **MAPE = 7.6**و **RMSE = 4.39**.وأكدت التحليلات الكمية والنوعية أن النقاط المفتاحية المتوقعة تطابقت بدرجة عالية مع القيم الحقيقية عبر صور التحقق المختلفة، مما يثبت قوة النموذج ودقته العالية في تحديد المعالم الوجهية.

الاستنتاج:

نجح الإطار الهجين المقترح في دمج مزايا عدة شبكات تلافيفية لتحقيق توازن بين **عمق التمثيل** و**تنوع السمات** و**الكفاءة الحاسوبية** .وقد حقق أداءً مرتفعاً وقدرة جيدة على التعميم في مهمة اكتشاف النقاط الوجهية، مما يجعله حلاً واعداً لتطبيقات متقدمة مثل **التعرف على العواطف** و**تحليل الوجوه** و**التفاعل بين الإنسان والحاسوب**.

الكلمات المفتاحية: اكتشاف النقاط الوجهية؛ التعلّم العميق؛ النموذج الهجين؛ الشبكات العصبية التلافيفية؛ الرؤية الحاسوبية.

ISSN: 2312-8135 | Print ISSN: 1992-0652
info@journalofbabylon.com | jub@itnet.uobabylon.edu.iq | www.journalofbabylon.com