Received: 28 September 2024 | Revised: 13 January 2025 | Accepted: 13 February 2025 | Published online: 1 April 2025

RESEARCH ARTICLE

Journal of Computational and Cognitive Engineering
2025, Vol. 4(4) 535–545
DOI: 10.47852/bonviewJCCE52024451

BON VIEW PUBLISHING

Ensemble Deep Learning Framework for Hybrid Facial Datasets Using Landmark Detection: State-of-the-Art Tools

Nandhini Krishnasamy[1,*], Nilima Zade[1], Dhruvi Khambholia[2], Rabinder Henry[3] and Aditya Gupte[4]

[1]*Symbiosis Institute of Technology-Pune, Symbiosis International (Deemed University), India*

[2]*Department of Artificial Intelligence and Machine Learning, Symbiosis International (Deemed University), India*

[3]*ATLAS SkillTech University, India*

[4]*Aviation Services Research Council, The Hong Kong Polytechnic University, China*

Abstract: Autonomous face emotion recognition (FER) with landmarks has become an important field of research for human–computer interaction. A significant achievement has been achieved through deep learning algorithms in recent times. Recognizing faces can be done using an end-to-end approach with deep learning techniques, which learns a mapping from raw pixels to the target label. In the field of emotional classification, the research community has extensively utilized 98 and 68 facial landmarks. In particular, pre-trained convolutional neural networks such as the residual network 50-layer network with the random sampler, Visual Geometry Group 16-layer network, and MobileNet including their ensemble versions of deep learning models are popular among researchers due to their ability to handle complex data. Researchers have mostly evaluated the model on a single dataset. A single dataset poses a challenge in developing a generalized model capable of capturing the full versatility of emotions. The key challenge in the dataset is that a single emotion is represented in multiple facial expressions with low-resolution images. This research study uses a combined dataset (CK+, KDEF, and FER-2013), which is more challenging than a single dataset. This research study offers a comprehensive analysis involving 68 and 98 landmarks with different FER deep models, examining how landmarking and different network architectures contribute to emotion recognition accuracy. This research study also considers the overfitting and class imbalance of the proposed ensemble model, which improves its performance by batch-wise feature extraction. Results show 78% accuracy with 98 landmarks and 75% with 68 landmarks. Overall, the model significantly reduces the gap between training and testing accuracy for both single and combined datasets.

Keywords: face emotion recognition (FER), face landmarks, weighted random sampler, ensemble model

1. Introduction

Given the ever-closer connection between people, comprehension of human feelings is essential in many areas, including computer–human communication, health care, advertisement, and show business. Real-time sentiment analysis is a key enabler for empathic and responsive technologies. One of the most effective approaches for this purpose entails using facial landmarks for emotion detection. For a time, the capacity of the human face to artistically convey the deepest emotions has been recognized. Even small movements in muscles can reveal a wealth of information about someone's feelings, ranging from surprise and excitement to anger and despair. The study presented focuses on human emotion recognition through facial landmarks. This work focuses on the underlying techniques that enable researchers to better understand emotions and equip technology with the capability to interact with people on a more personal level. From the expressions on someone's face, their mental state can be understood. This knowledge finds use in such domains as psychology, marketing, and human–computer interaction (HCI) [1]. A given method that looks promising includes analyzing landmarks that are distinctive body points in each individual. In turn, this helps to better understand the feelings of the individual. These landmarks, located near the corners of the eyes, nose, and mouth, offer useful hints on facial expressions and emotions [2]. The main analytical difference lies in the complete comparison of the effectiveness of facial markers 68 and 98 in emotion recognition. There is still a lot of scope in research on how increasing landmark granularity affects model accuracy under different conditions, especially on single and combined datasets.

In addition, with the rapid progress of technology, there has been an abundance of high-quality cameras and improved computing capabilities, resulting in immense progress in face recognition research. Though more research into machine learning techniques is still required, the interest in dynamic models is rising, as these models move away from static, pre-recorded dialogue toward online systems that are sensitive to the subtleties of the user response.

***Corresponding author:** Nandhini Krishnasamy, Symbiosis Institute of Technology-Pune, Symbiosis International (Deemed University), India. Email: nandhini.k@sitpune.edu.in

For that reason, new expressions and algorithms have been introduced, which have the potential to recognize the unique facial traits in different manners of human emotions. However, the trip is just beginning in this challenging and complex task. The challenges include better granulation of emotion categorization, adaptability to diverse cultural expressions, and real-time processing optimization [3]. Similarly, issues related to privacy and prejudice are important when employing such technologies that become ever more widespread. Datasets in face emotion recognition (FER) can introduce biases due to the underrepresentation of certain demographics or imbalanced emotional categories. Such biases may result in models that perform inconsistently across diverse populations or fail to generalize effectively. These biases can be mitigated by using ensemble models, augmentation techniques, and the use of diverse datasets; hence, we have used the combined dataset approach. The major contributions to this research study, crafted by considering the above challenges in a real-time scenario, are as follows: (1) Developed a robust methodology ensemble deep learning framework to identify emotions with low-resolution images. (2) Adapted and compared the crucial steps of 68 and 98 landmarks for facial emotion recognition. (3) Utilized state-of-the-art tools by enabling various deep learning frameworks for facial emotion recognition.

The article is organized as follows. In the first section, the Introduction consists of a statement of purpose and objectives with the research gap. In the second section, the background work includes related work and a summary of the literature review. The third section on methodology includes the methods used and a detailed discussion of the dataset and its experimental parameters. The fourth result section includes experimental results and discussion, followed by the fifth section as the conclusion and future scope.

2. Background Work

Key features of the face, called facial markers, are used to convey emotions and to recognize and to evaluate facial expressions. By facilitating the extraction of important features from facial images, these landmarks improve the accuracy of emotion recognition algorithms. Many existing models and techniques have been established to identify and use these markers effectively. This review covers the major contributions in this field, focusing on works published in the last decade. In the literature, it has been identified that face detection, landmark localization, and pose estimation are distinct methodological frameworks that have been implemented. The unified tree-structured models for landmark localization, face detection, and pose estimation are framed to address the global mixtures in a viewpoint topological capture, and facial landmarks in every part achieved an improved result in FER. Gao et al. [4] introduce an ensemble approach that enhances facial landmark detection, especially in challenging conditions like extreme angles and heavy occlusions. Tests on complex datasets like Menpo and Caltech Occluded Faces in the Wild show that the proposed method improves accuracy and performs well even in difficult scenarios. The work suggested the potential gap between academic and commercial research due to the unavailability of vast, diverse datasets. Du et al. [5] present a lightweight and fast facial landmark detection model based on an improved YOLOv5n with an attention mechanism, enhancing accuracy and speed. Deployed on a Jetson Nano, the model effectively identifies key facial points in real time, achieving high recall and accuracy rates that make it suitable for edge devices. Teoh et al. [6] explore the design of a face recognition system using deep learning with OpenCV in Python, highlighting its reliability and high accuracy. The system effectively identifies individuals based on facial features, making it secure and practical for applications like phone unlocking, security, and criminal identification. Experimental results support the system's accuracy and robustness. However, the effectiveness of the proposed method needs to be tested on combined datasets. The 68-point visage model from the dlib library was released in 2009 [7] and quickly gained popularity as a standard in the industry. This model provided a reliable tool for the FER task due to its accuracy in critical areas around the mouth, nose, eyes, and chin. Vachmanus et al. [8] proposed a deep alignment network where a set of regression trees is used to solve the problem of angle disturbances and angle changes. Their method performed well when it came to identifying markers with different shapes and partial curvatures. However, new strategies need to be investigated for the variability of datasets. LRR-Net is used to optimize deep learning models [9]. Wider Facial Landmarks in-the-wild (WFLW) dataset—which had a 98-point facial landmark model—was provided by Wu et al. [10]. The capabilities of the FER system were enhanced by adding additional points that provided an extended sample, which helped capture more tiny facial expressions. The authors discussed the challenge that few facial expressions cannot be categorized as unique emotions, which leads to increased misclassification. The work of Mollahosseini et al. [11] examined the recognition of multiple emotions by mixing facial features with physical and auditory cues. This holistic approach improved the accuracy of emotion recognition systems. However, the generalization of such an approach can be a challenge.

Bulat et al. [12] proposed an Adversarial Graph Representation Adaptation framework with a residual network 50-layer network (ResNet50) as the backbone. The author has developed a real-time facial mark recognition system, adapting the model for live applications and enhancing its suitability for communication systems and real-time monitoring. The proposed method's effectiveness needs to be tested on combined datasets. A lightweight facial mark recognition model intended for mobile devices was presented by Chen et al. [13]. Their work has enabled better emotion recognition for portable devices by addressing the computational limitations of mobile platforms. However, extensive research is needed to detect emotions correctly in pose-invariant cases. Starting in 2022, the work presented by Gao et al. [14] included approaches to improving the recognition of facial symbols. The approach enhanced the artist's ability to focus on relevant facial regions, thus increasing perceptual accuracy and marker recognition. However, the robustness of the model needs to be validated on variable and combined datasets. Facial expression recognition has improved greatly due to the development of facial landmark recognition techniques. From simple geometric techniques to complex deep learning models, each advance has addressed a specific issue and increased the flexibility and accuracy of emotion recognition in general. Subsequent research is expected to further develop these models by increasing their performance for real-time processing on data-rich platforms. The AffectNet dataset, a massive resource for education emotion detection algorithms, has been provided by Ahmed et al. [15]. They gave an instance of the way facial landmarks may increase accuracy.

It is observed that most studies widely use smoothing, filters, and dimensionality reduction techniques such as principal component analysis, optical flow, and local binary pattern for FER. The issue of resolving the complexity related to pose variants/uncontrolled environments in FER is difficult and leads to a lack of accuracy in conventional FER systems [16]. Facial localization is one of the techniques that can be included in which it uses landmarks, and extraction will be more effective on fiducial features such as the eyebrows, lips, nose, and eyes. Adapting traditional detection methods for corners and edges, which can be evaluated using Euclidean distance, will enable the generation of new

features. These newly generated feature vectors can be trained with a multilayer perceptron network to classify the inherent nature of face expressions [17]. Achieving the improved classification result in the limited computed environment for FER, especially in real-time vision systems, required a lightweight convolutional neural network (CNN). Multi-task cascaded neural networks bring the facial coordinates to the classification model to detect various facial features, which also require fewer memory resources [18]. Another approach to detecting real-time streaming FER by enabling transfer learning in CNN, especially on the fully connected layers, shows significant improvement in accuracy [19]. In Ghimire and Lee [20], researchers proposed a bagging ensemble with a histogram of orientation gradient feature extraction for FER. Models have been tested on the Extended Cohn-Kanade (CK+) and Japanese female facial expression (JAFFE) datasets individually. Results show significant improvement; however, the robustness of the model needs to be tested on a combined dataset.

Recent advancements in landmark detection for FER uncover potential opportunities and benefits as the industry unveils different applications. Face landmarks have gained popularity due to their significant role in healthcare and biomedical applications, facial recognition, facial expression analysis, driver monitoring systems, augmented reality and virtual reality applications, etc. In particular, video-based assessment of Parkinson's disease with facial landmarks has a great health analysis impact. Significant challenges in landmark detection occur in unconstrained environments, such as occlusion, lighting conditions, backgrounds, and varying poses [21, 22]. From the study, a potential research gap identified is the poor performance of FER models due to uncontrolled environmental factors, such as lighting conditions, camera angles, and occlusions. Landmarks play a crucial role in improving FER accuracy, typically using 68 or 98 key points. FER is an essential component of HCI, and it aims at enabling machines to perceive the emotional states felt by humans. Although various models have been proposed using machine learning or deep learning, few studies have developed the method of fusing multiple datasets to improve FER performance. In this research paper, we looked at the performance of concatenating a pair of datasets for better results in FER. Even if different datasets have their unique features and biases, the result of combining multiple ones is a more complete training resource. Such fusion can help to avoid overfitting and enhance the generalization ability of FER models. This could include using datasets with diverse lighting conditions, ethnic backgrounds, and emotional expressions, which would give more robustness to the models in real-world applications.

3. Methodology

Despite progress, FER systems often miss subtle facial emotions [23]. Existing techniques struggle with real-world challenges like lighting changes and occlusions. They also have difficulty with varying facial features, especially less obvious emotions. Traditional models—which use 68-point landmarks—lack the granularity that is needed for accurate emotion recognition. This gap is addressed by advanced 98-point markers integrated with state-of-the-art deep learning algorithms such as ResNet50, Visual Geometry Group 16-layer network (VGG16), and MobileNet and optimized using ensemble methods and hyperparameter tweaking. These methods not only increase detection accuracy in the combined dataset but also provide a more robust and generalizable model. It can work reliably in many situations.

3.1. Proposed framework

The acquisition of input data, including images from several facial expression datasets such as CK+, Karolinska Directed Emotional Faces (KDEF), and FER-2013, is the first step in the proposed procedure. The facial expressions of the datasets of these types are important for accurate emotion recognition. Pre-processing is applied to the images before models are employed. This includes "normalization," which is the process of bringing pixel values into a specified range. Rescaling is used to achieve normalization. Additionally, "resizing" ensures that each image has the same dimensions. Figure 1 illustrates the steps involved in the framework.

Facial landmarks: Facial landmarks, which are extraordinary features of the face, are now a critical element of emotion recognition structures. These locations act as fulcrums for the monitoring and evaluation of facial expressions. Figure 2(a) [24] and (b) [10] depict noteworthy contributions in this area. After pre-processing images, the next important step is marker recognition. It uses marker

Figure 1
Block diagram of the proposed methodology framework

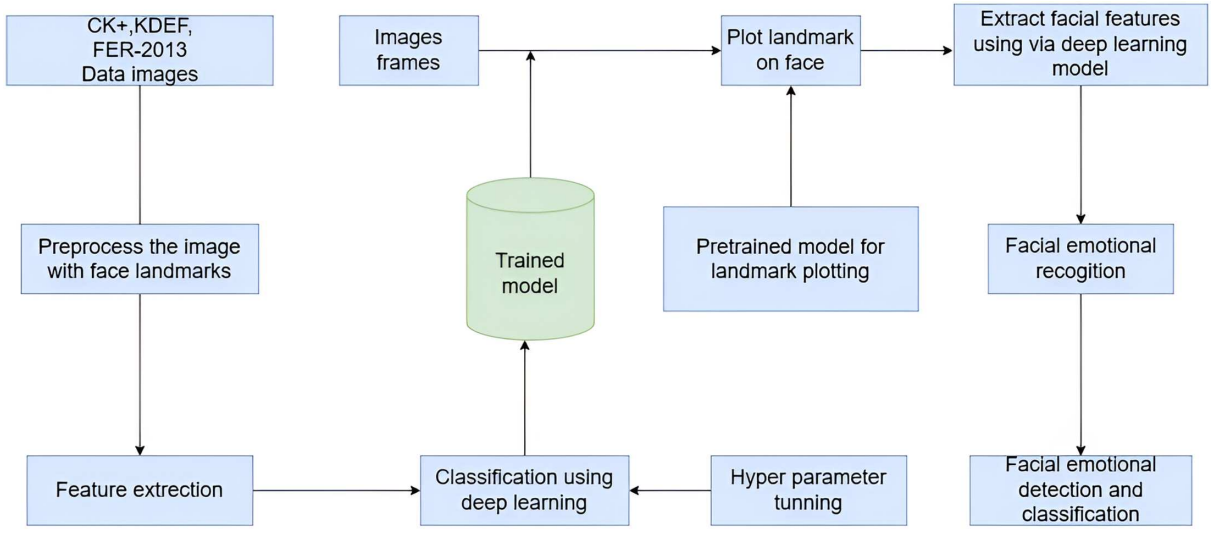
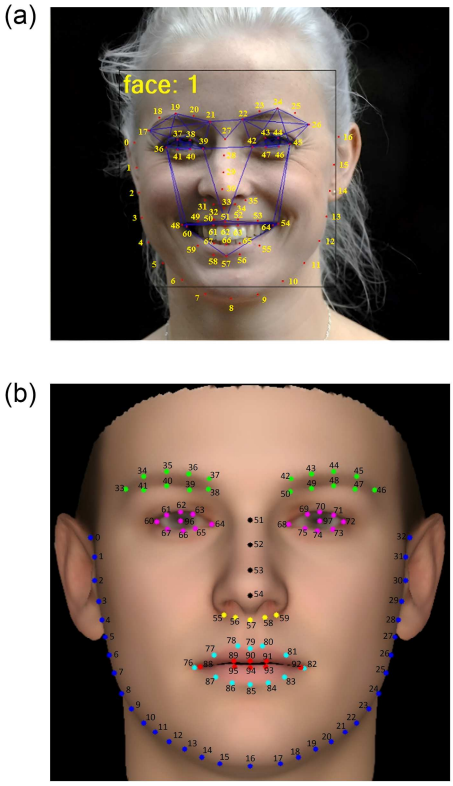


Figure 2
(a) 68 landmark positions and (b) 98 landmark positions



recognition techniques to identify important points in the face. There are two main types: (1) The 68-Point dlib: This model displays 68 points on the face to capture key features such as the mouth, nose, eyes, eyebrows, and jaw. (2) The 98-point WFLW: This model provides a robust map with 98 points, which facilitates fine-grained feature extraction and improves comment recognition.

After identifying key points on the face, the system collects relevant attributes and assigns them to multiple deep learning models. These examples include CNN, which can extract facial features like the shape and angles of the nose, eyes, lips, and cheeks during different expressions from images. ResNet50 is a residual network with a deep architecture that facilitates the learning of complex models. VGG16 has a 16-layer deep mesh well known for its straightforward and efficient image classification applications. MobileNet is lightweight graphics designed for mobile-edge devices to balance accuracy and performance. By using the combined data, each model is trained to predict the emotions associated with the retrieved attributes. To reduce the loss function and increase the prediction accuracy, the models iterate over the data during the training process, which spans multiple periods. The performance of each model is optimized by hyperparameter adjustment.

The framework uses a group approach to exploit the unique potential of each sample. It combines the results of CNN, ResNet50, VGG16, and MobileNet models. They can be used in various ways. For example, weighted averaging combines predictions from each model, assigning weights based on their performance to reach the final prediction, while majority voting is the decision model's final output, which is the most shared prediction among all models. The ensemble model is further optimized using optimization techniques such as cross-validation and hyperparameter tweaking to ensure high accuracy and smoothness. The classification of emotions based on the results of the group model is the final stage of the process. The detected facial markers and their corresponding characteristics are divided into several emotion categories such as surprise, happiness, sadness, and anger, after which the results are shown. The performance metrics are used for each model, which include accuracy, precision, recall, and $F1$-score. Confusion mMatrix and accuracy-loss graphs are used to compare model performance, and their visualization shown in tables and graphs played a role in the study.

3.1.1. Dataset description

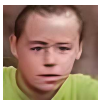
CK+: Cohn-Kanade extended (CK+) dataset contains 920 individual facial expressions. Data has already been reshaped to 48 × 48 pixels, in grayscale format, and face cropped using haarcascade frontal face default. It is an open-source dataset [25].

KDEF: KDEF dataset reshaped into 512 × 512 super-resolution and colorized. It is an open-source dataset. It contains 4900 images [26].

FER-2013: This dataset consists of grayscale images of faces in 48 × 48 pixels. One of the seven categories best describes the emotion shown in the facial expression [27].

The combined data is shown in Table 1. By combining data from images of varying resolutions, a more versatile model can be trained. Such a model is capable of handling a wider range of inputs. This approach not only enables the training of more generalized models but also allows for a more diverse dataset, enhancing

Table 1
Training and testing dataset

Class labels	Emotion	Train	Test	Sample images
0	Anger	4325	400	
1	Contempt	112	18	
2	Disgust	682	113	
3	Fear	2961	493	
4	Happiness	8048	1001	
5	Neutrality	4348	724	
6	Sadness	4803	600	
7	Surprise	3727	499	

the overall robustness and adaptability of the models. The combination of CK+, KDEF, and FER-2013 provides a robust and scalable dataset that can greatly improve the construction and evaluation of facial expression recognition systems.

Each dataset has advantages and disadvantages. For example, KDEF provides different faces, CK+ provides more detailed sequences, and FER-2013 has a larger sample size. Their combination assures a thorough examination of all emotions and their differences. Such a combined dataset is more robust and reflects real-world conditions due to facial expressions, locations, lighting conditions, and the increased number of subjects.

3.2. Architectural design

CNN: CNNs are often used for complex applications such as object detection [28], face mask detection [29], and facial landmark detection [30, 31]. The network can process facial images or landmarks to identify spatial patterns and features, making them essential for accurately locating key points on a face. In this model, the input image size is 64, and the ReLu activation function is used along with the size of 32 filters in every layer. There are four layers of convolutional and max pooling to filter the features. It is a categorical-based model, for loss count categorical cross entropy is used with an "Adam" optimizer.

ResNet50: This model is deeply connected, and 50 layers are there in the model so that the model can be trained in many stages. Its architecture gives a more precise result than simple neural network models. The reason behind this is 50 layers of big architecture. "Adam" optimizer is used for optimization and categorical cross entropy is used to show loss because there are multiple classes for classification. Figure 3 [30] represents the basic architecture of the ResNet50 model. The weighted random sampler algorithm is a technique used for handling imbalanced datasets during the training of the ResNet50 model. It is more effective when dealing with datasets where certain classes are underrepresented. The weighted random sampler assigns different weights to each sample in the dataset, influencing the probability of each sample being selected during the training process. The weighted random sampler is a valuable tool for addressing class imbalance during training. It helps prevent the model from being biased toward the majority class by ensuring that the model sees a balanced representation of all classes throughout the training process.

VGG 16: The pre-trained weights for VGG16 on large image datasets (like ImageNet) are available, enabling transfer learning for various computer vision tasks. VGG16 has a relatively large number of parameters, making it computationally expensive and memory-intensive. The extensive use of small convolutional filters contributes to the higher number of parameters. VGG16 has a relatively large number of parameters, making it computationally expensive and memory-intensive. Figure 4 [31] indicates the basic architectural model of the VGG16.

MobileNet: MobileNet uses a specific type of convolutional layer called depth-wise separable convolutions to reduce computational complexity. Depth-wise separable convolutions consist of two steps: depth-wise convolutions, which filter the input channels separately, and pointwise convolutions, which combine the outputs from the depth-wise step using 1×1 convolutions. This architecture significantly reduces the number of parameters and computations compared to traditional convolutional layers. MobileNet is commonly used for tasks like image classification, object detection, and

Figure 3
The basic architecture of ResNet 50 model

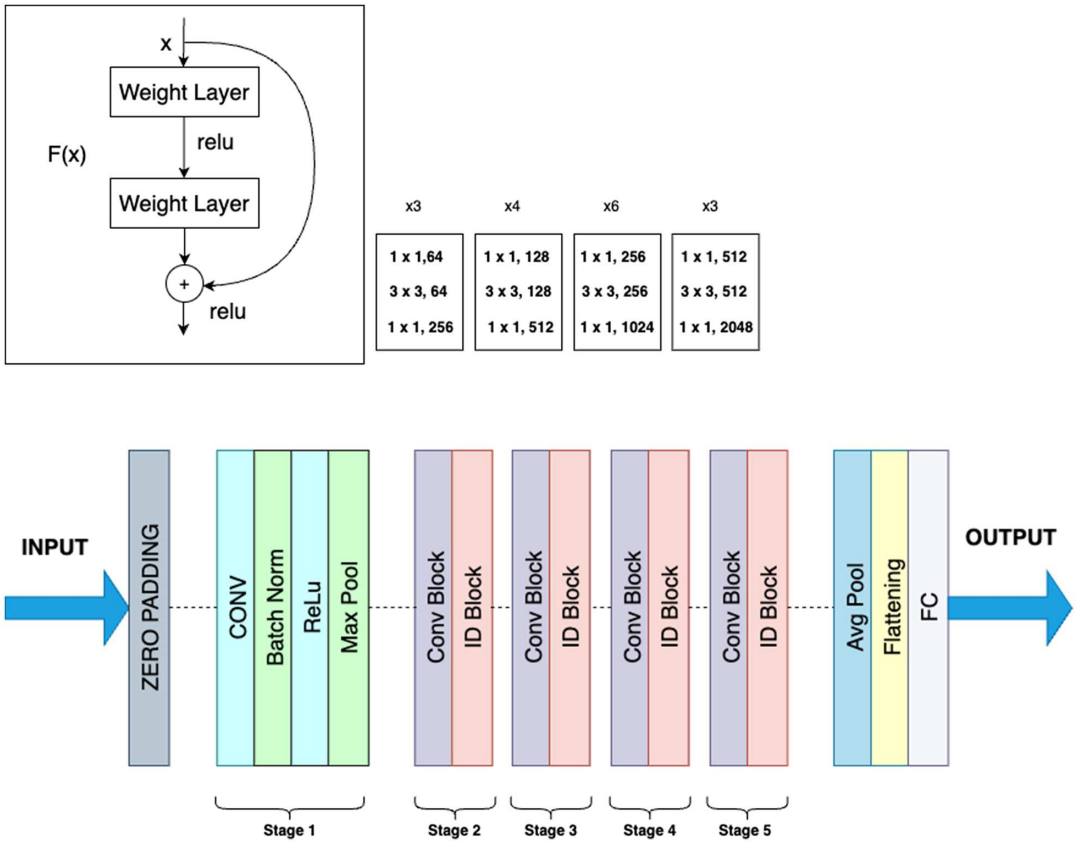


Figure 4
The basic architecture of VGG16 model

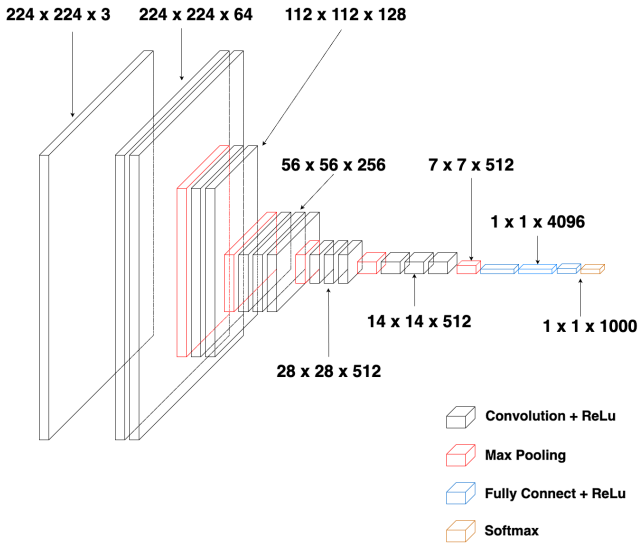


image segmentation in scenarios where computational resources are limited. Table 2 represents the implementation details of the deep learning models.

Proposed ensemble ResNet50 and VGG16: An ensemble method is a machine learning technique that combines the predictions of multiple individual models to improve overall performance, especially when the individual models are diverse and complementary to each other. Instead of relying on a single model, ensemble methods leverage the diversity of multiple models to make more accurate predictions. Figure 5 indicates the architectural representation of the ensemble model. An optimized ensemble method has been proposed for face emotion detection using 68 and 98 facial landmarks. To combine the ResNet50 and VGG16 models, fine-tuning was applied to each model, along with techniques such as data augmentation, weighted random sampling, and hyperparameter tuning. These methods enhanced the ensemble's accuracy and robustness in recognizing emotions by gradient boosting. Training and testing accuracy were the primary parameters in these experiments.

The data has been split into training and testing subsets. After normalization, deep models were employed. CNN, ResNet50, VGG16, MobileNet, Ensemble of VGG16, and ResNet50 have been employed on the combined datasets. A total of 50 epochs are set for all models in the training and testing phase and accuracy is observed. In the study, if the accuracy remains constant across consecutive iterations, then the early stopping method is implemented. The parameters like batch size, image input size, accuracy, and number of epochs are tuned to get the best accuracy for the employed models. The models are compared and evaluated to justify the necessity of this proposed model.

4. Results and Discussion

Although all the models performed well, hyperparameter optimization of VGG16 and ResNet50 ensembles yielded the highest accuracy in the dataset. Standard CNN and MobileNet achieved good baseline results but were outperformed by deeper models such as ResNet50 and VGG16. The random sampler employed in ResNet50 had better handling of class imbalance, although it was surpassed by the optimized ensemble, which demonstrated the advantages of mixing different models and further tuning them to make robust detection of emotions in human faces.

Table 3 represents the results of training and testing accuracy on 68 landmarks and 98 landmarks in the combined dataset. These observations indicate that the 68 landmarks model was weaker than 98 landmarks. The difference between the performance of 68 and 98 landmarks was marginal in this study as combined datasets consist of low-resolution, gray, and colored images. The MobileNet model achieved training and testing accuracy of 72% and 61%, which contains a minimal architecture where computational resources are limited, so it cannot be considered for an uncontrolled environment. VGG16 and ResNet50 demonstrate better performance when compared with the CNN model. The proposed ensemble model outperformed in the testing environment when compared to all other models. When the performance of the ResNet50 is observed, it is found that the training and testing accuracy deviates by ~20%. Hence, the proposed ensemble model is tested and observed that the deviation in accuracy has been reduced to ~10%, which indicates that it can perform better in an uncontrolled environment. Table 4 represents the performance of the ensemble model on single and combined datasets. It is observed that the ensemble model

Table 2
Summary of deep learning model implementation

Implementation details	CNN	ResNet50	VGG16	MobileNet
Convolutional Layers	5 layers with different filter sizes (32, 64) and "Sigmoid/Softmax" activation	48 layers with different filter sizes (32, 64, 128, 256) and "Sigmoid/Softmax" activation	13 layers with different filter sizes (32, 64, 128, 256) and "Sigmoid/Softmax" activation	10 layers with different filter sizes (16, 32, 64) and "Sigmoid/Softmax" activation
Pooling Layer	3 Max pooling layer	1 Max pooling layer	5 Max pooling layers	1 Global Avg pooling layer
Flatten Layer	NA	NA	1	1
Dense Layer	Fully connected with Dense Layers with 256 units		3 with 256 units	1 with 1024 units
Loss Function	Categorical cross entropy for all models			
Batch Size	32	32	64	128
Optimizer	ADAM for all models			
Steps per epoch	907	907	363	226

Figure 5
Ensemble method of VGG16 and ResNet50 proposed block diagram

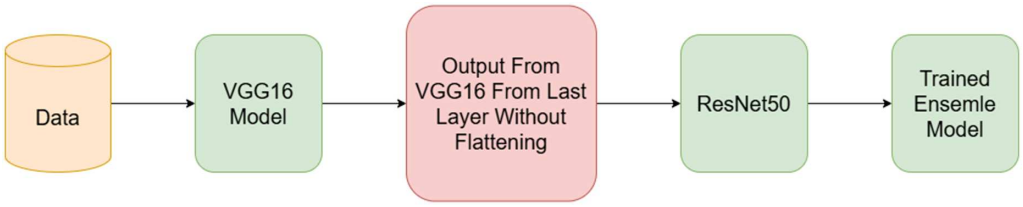


Table 3
Training and testing accuracy for 68 and 98 landmarks on combined dataset

Model	68 landmarks		98 landmarks	
	Training accuracy	Testing accuracy	Training accuracy	Testing accuracy
CNN	55%	65%	50%	63%
ResNet50	85%	64%	82%	64%
VGG16	70%	62%	73%	63%
MobileNet	73%	60%	72%	61%
Ensemble	75%	66%	78%	67%

performed significantly well compared to other models, especially with the FER-2013 and KDEF datasets although the datasets are imbalanced. The individual results of the FER-2013 and KDEF imbalance datasets show that it affects the performance of the combined dataset compared to the individual Ck+ dataset.

Table 4
Results ensemble model with the single and combined dataset

Dataset	98 landmarks	
	Training accuracy	Testing accuracy
Ck+	89.78%	89.15%
FER- 2013	41.75%	44.56%
KDEF	62.7 %	65.3%
KDEF + FER-2013 + CK+	78%	67%

Table 5 shows the comparison of published research studies in which the models are tested on a single dataset. Khan [17] demonstrates that training and testing accuracy deviates by more than ~40% in the CNN model on the FER-2013 dataset. The proposed work exhibits only up to 10% variation in the training and testing phase, suggesting a stable model. Duncan et al. [19] indicate that the ensemble model demonstrates significant performance on individual datasets. These observations also demonstrate that single data can work well, but there is a lack of data diversity, so generalization is a necessity for open-use cases. It has been observed that different techniques applied to the same dataset often result in varying accuracy. To address this, combining the datasets approach with an ensemble method is employed. This approach helps to create a more generalized model that can adapt to different environments, such as lighting conditions, and facial features, ensuring more accurate detection results.

Table 6 shows that landmarks play a significant role in facial emotion recognition. When fewer landmarks are used with an imbalanced dataset, the individual performance for classes such as anger, contempt, disgust, and neutrality shows poor accuracy. However, by increasing the number of landmarks to 98, the results improve as the precise points on the face model extract more accurate features, leading to better outcomes. Figures 6 and 7 represent the confusion matrix for the ensemble model using 68 and 98 landmarks. It can be observed that a few emotions such as happiness, neutrality, sadness, and surprise have more chances to misclassify among themselves than other emotions. This suggests that more landmarks need to be explored to classify such emotions.

Table 6 shows improvements in classes 1, 2, 4, and 5 due to the transition of 68 and 98 landmarks, especially with evaluation parameters precision and recall. Table 7 shows the ablation analysis. The proposed model benefits due to an increase in the landmarks, but there is a negative impact on classes 3 and 5. This may occur due to model complexity or overfitting. Finally, it is observed that 98 landmarks may introduce additional computational complexity, which leads to a slow training process, but the improvements in precision and recall trade-offs are justified.

In summary, the combined dataset has more complexity than any single dataset in the tested environment. Comparative analysis underscores the importance of addressing class imbalances for improved model performance, especially in the context of ResNet50. The weighted random sampling approach appears beneficial, but its impact on overfitting should be carefully managed for optimal results in real-world applications. Based on the presented results, the choice between models depends on the specific requirements and constraints of the application. The CNN model, although less complex, exhibits lower overall accuracy, suggesting a trade-off between model complexity and performance. When a validation image was given to the trained model, true positives were achieved for both 68 and 98 landmarks. However, when given to individual models like CNN, MobileNet, and ResNet50, some failed to

Table 5
Comparison of training and testing accuracy for 68 and 98 landmarks with published work

Ref No.	Dataset	Model	Training accuracy	Testing accuracy
[1]	CK+	Fuzzy logic approach	83.2%	71.5%
[3]	AffectNet	CNN	69.3%	–
[18]	FER-2013	CNN, ResNet50	CNN 99.70%, ResNet50 98.87%	CNN 58.90%, ResNet50 57.48%
[20]	CK+, JAFFE	Ensemble learning model (ELM) bagging ensemble	–	CK+ (97.3%), JAFFE (94.37%)
Proposed work	KDEF + FER 2013 + CK+	Ensemble (with hyperparameter tuning)	75% (with 68 landmarks), 78% (with 98 landmarks)	66% (with 68 landmarks), 67% (with 98 landmarks)

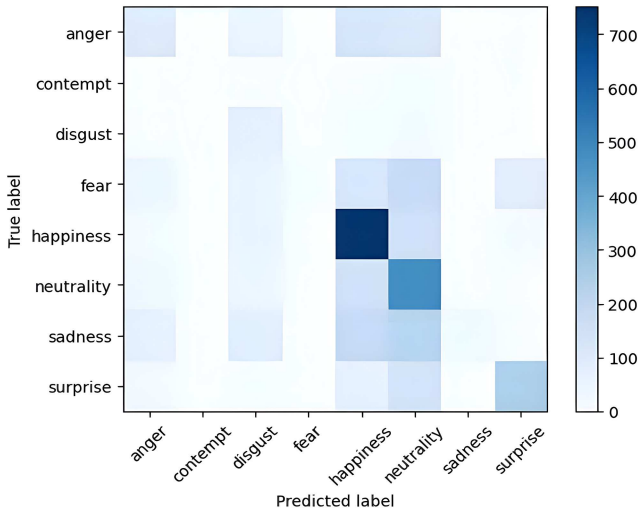
Table 6
Classification report of ensemble method for 68 and 98 landmarks

	68 landmarks			98 landmarks		
Class	Precision	Recall	*F*1-sore	Precision	Recall	*F*1-sore
0	0.35	0.24	0.28	0.55	0.62	0.58
1	0.00	0.00	0.00	1.00	0.17	0.29
2	0.19	0.63	0.30	0.83	0.59	0.69
3	0.64	0.02	0.04	0.63	0.35	0.45
4	0.52	0.75	0.62	0.75	0.93	0.83
5	0.36	0.67	0.47	0.67	0.54	0.60
6	0.82	0.05	0.09	0.60	0.62	0.61
7	0.66	0.50	0.57	0.68	0.80	0.74

Table 7
Ablation study analysis for proposed model with 68 and 98 landmarks for FER dataset

Study component	68 landmarks	98 landmarks	Difference/impact
Model Architecture	Proposed model with 68 landmarks for FER	Proposed model with 98 landmarks for FER	Additional landmarks led to higher granularity
Data Pre-processing	Standard face alignment steps	Standard alignment + 98 landmark adjustments	Minimal impact on pre-processing complexity
Precision (Average)	0.44	0.72	+0.28 (98 landmarks significantly improved precision)
Recall (Average)	0.36	0.58	+0.22 (better detection of expressions)
*F*1-Score (Average)	0.3	0.6	+0.30 (overall balanced performance improvement)
Training Time	Lower due to fewer features	Slightly higher with 98 landmarks	Training time increased slightly but was acceptable, given the performance boost
Inference Efficiency	Faster due to fewer features	Marginally slower with more landmarks	Impact manageable compared to accuracy gains

Figure 6
Confusion matrix of ensemble model of 68 landmarks



predict true values. The experimental study of comparative analysis shows that the proposed ensemble model with other models including CNN, ResNet50, VGG16, and MobileNet demonstrates better performance.

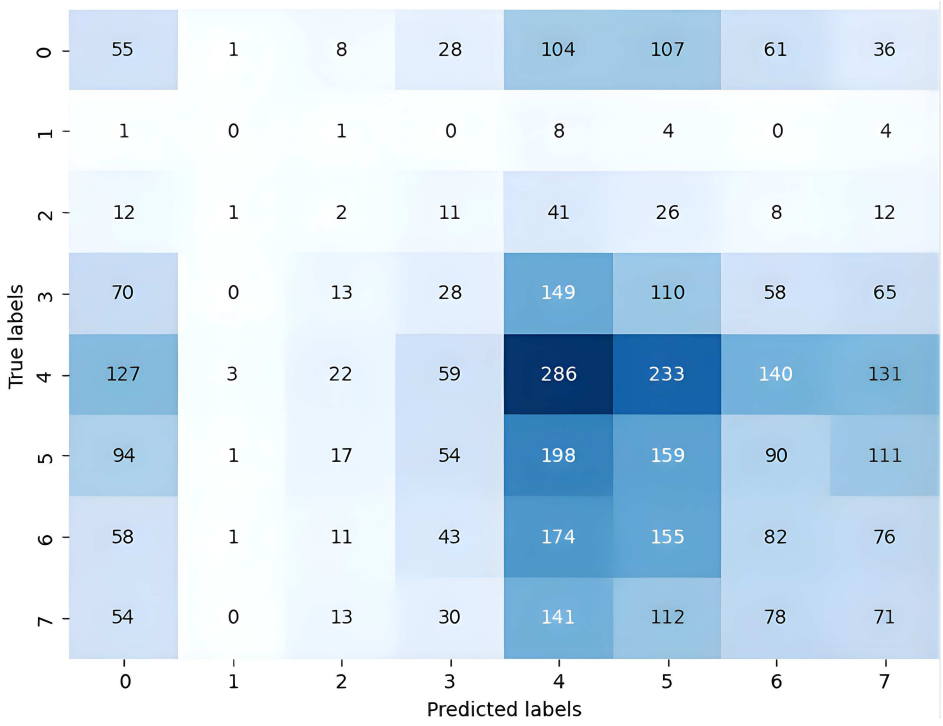
FER powered by an ensemble model can significantly improve HCI. They can provide valuable insights to therapists or serve as tools in self-monitoring apps, contributing to timely interventions and personalized mental health care. By analyzing customers' facial expressions during interactions, businesses can refine product placement, marketing strategies, and customer service approaches. For online courses, emotion-aware systems can provide feedback to instructors about learner focus and emotional states, enhancing learning outcomes. Ensemble models can improve the accuracy and reliability of such systems, reducing false positives and enhancing public safety. In the entertainment industry, such models can be useful for video games and interactive media that can adjust storylines or difficulty levels based on players' emotional feedback. Such systems can also provide insights for human resource policies and strategies to enhance productivity by identifying stress levels and job satisfaction. While the societal benefits of emotion recognition are significant, it also raises ethical concerns regarding privacy, consent, and potential misuse. Industries deploying these systems must adopt strict regulations, transparency, and opt-in mechanisms to ensure that individuals rights are respected.

5. Conclusion

This research study used 68 and 98 landmarks to investigate the use of different deep learning architectures for FER. These models include VGG16, ResNet50, MobileNet, CNN, and proposed ensemble models. The objective of the study was to build a robust and comprehensive model that can reliably identify facial expressions in uncontrolled environments using aggregated datasets from CK+, KDEF, and FER-2013. Model performance was further enhanced by the inclusion of hyperparameter tweaking, which ensured that each architecture was set to deliver maximum results on these datasets. The results show that compared to a single model, grouping methods—in particular, the combination of VGG16 and ResNet50—significantly increase the accuracy and reliability of emotion recognition. Combining the advantages of both proposals, the ensemble method captures finer spatial information and better handles complex interactions. The study also highlighted the importance of datasets for robust emotion recognition algorithms.

The scope of future work includes several directions for continued learning and improvement. To improve the model's ability, key attributes that are important for emotion recognition need to be

Figure 7
Confusion matrix of ensemble model of 98 landmarks



focused on in the future. Transformer-based topologies and sophisticated conceptual techniques can be added. To further improve the performance of the imbalance dataset, the generative AI approaches can be incorporated in the future. The dataset can be expanded to include more diverse real-world scenarios, such as dynamic expressions, and changing environmental conditions. Combining more information—for example, facial expressions, voice, and physiological signals—is needed to develop detailed and accurate perceptual recognition systems. Tiny U-NET in larger U-NET facilitates representation learning for multilevel and multiscale objects developed for infrared object detection. This may be incorporated in the future to extract fine-grained information on subtle features of facial regions.

Acknowledgment

The authors are grateful to the authorities of Symbiosis Institute of Technology Pune, Symbiosis International (Deemed University), Pune, India, for their support.

Funding Support

The work is supported by the Research Support Fund of Symbiosis International University.

Conflicts of Interest

The authors declare that they have no conflicts of interest to this work.

Data Availability Statement

The CK+ data that support the findings of this study are openly available in Kaggle at https://www.kaggle.com/datasets/shuvoalok/ck-dataset. The KDEF data that support the findings of this study are openly available in Kaggle at https://www.kaggle.com/datasets/tom99763/testtt. The FER2013 data that support the findings of this study are openly available in Kaggle at https://www.kaggle.com/datasets/msambare/fer2013.

Author Contribution Statement

Nandhini Krishnasamy: Conceptualization, Methodology, Software, Formal analysis, Investigation, Data curation, Writing – original draft, Visualization, Supervision, Funding acquisition. **Nilima Zade:** Conceptualization, Methodology, Validation, Formal analysis, Investigation, Data curation, Writing – original draft, Visualization, Supervision, Project administration. **Dhruvi Khambholia:** Conceptualization, Methodology, Software, Formal analysis, Investigation, Data curation, Writing – original draft, Visualization. **Rabinder Henry:** Validation, Formal analysis, Data curation, Writing – review & editing, Visualization, Supervision, Project administration, Funding acquisition. **Aditya Gupte:** Validation, Formal analysis, Data curation, Writing – review & editing, Visualization, Supervision, Project administration.

References

[1] Bahreini, K., van der Vegt, W., & Westera, W. (2019). A fuzzy logic approach to reliable real-time recognition of facial emotions. *Multimedia Tools and Applications*, *78*, 18943–18966. https://doi.org/10.1007/s11042-019-7250-z

[2] Zhao, Z., Liu, Q., & Zhou, F. (2021). Robust lightweight facial expression recognition network with label distribution training. *Proceedings of the AAAI Conference on Artificial Intelligence, 35*(4), 3510–3519. https://doi.org/10.1609/aaai.v35i4.16465

[3] Mukhiddinov, M., Djuraev, O., Akhmedov, F., Mukhamadiyev, A., & Cho, J. (2023). Masked face emotion recognition based on facial landmarks and deep learning approaches for visually impaired people. *Sensors*, *23*(3), 1080. https://doi.org/10.3390/s23031080

[4] Gao, H., Yang, X., Hu, Y., Xu, H., Liang, Z., Wang, B., ..., & Hu, S. (2024). Facial landmark detection algorithm in complex scenes. In *9th International Conference on Control and Robotics Engineering*, 352–358. https://doi.org/10.1109/ICCRE61448.2024.10589885

[5] Du, H., Zhang, X., Song, H., Lin, Y., & Li, L. (2023). Facial landmark detection based on improved YOLOv5 and edge computing. In *8th International Conference on Intelligent Computing and Signal Processing*, 1359–1363. https://doi.org/10.1109/ICSP58490.2023.10248697

[6] Teoh, K. H., Ismail, R. C., Naziri, S. Z. M., Hussin, R., Isa, M. N. M., & Basir, M. S. S. M. (2021). Face recognition and identification using deep learning approach. *Journal of Physics: Conference Series, 1755*(1), 012006. https://doi.org/10.1088/1742-6596/1755/1/012006

[7] King, D. E. (2009). Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, *10*, 1755–1758.

[8] Vachmanus, S., Phinklao, N., Phongsarnariyakul, N., Plongcharoen, T., Hotta, S., & Tuarob, S. (2024). Automating manga character analysis: A robust deep vision-transformer approach to facial landmark detection. *IEEE Access*, *12*, 131284–131295. https://doi.org/10.1109/ACCESS.2024.3459419

[9] Li, C., Zhang, B., Hong, D., Yao, J., & Chanussot, J. (2023). LRR-Net: An interpretable deep unfolding network for hyperspectral anomaly detection. *IEEE Transactions on Geoscience and Remote Sensing*, *61*, 5513412. https://doi.org/10.1109/TGRS.2023.3279834

[10] Wu, W., Qian, C., Yang, S., Wang, Q., Cai, Y., & Zhou, Q. (2018). Look at boundary: A boundary-aware face alignment algorithm. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2129–2138.

[11] Mollahosseini, A., Hasani, B., & Mahoor, M. H. (2019). AffectNet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, *10*(1), 18–31. https://doi.org/10.1109/TAFFC.2017.2740923

[12] Bulat, A., Sanchez, E., & Tzimiropoulos, G. (2021). Subpixel heatmap regression for facial landmark localization. In *32nd British Machine Vision Conference*, 1–15.

[13] Chen, T., Pu, T., Wu, H., Xie, Y., Liu, L., & Lin, L. (2022). Cross-domain facial expression recognition: A unified evaluation benchmark and adversarial graph learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *44*(12), 9887–9903. https://doi.org/10.1109/TPAMI.2021.3131222

[14] Gao, P., Lu, K., Xue, J., Lyu, J., & Shao, L. (2023). A facial landmark detection method based on deep knowledge transfer. *IEEE Transactions on Neural Networks and Learning Systems*, *34*(3), 1342–1353. https://doi.org/10.1109/TNNLS.2021.3105247

[15] Ahmed, N., Al Aghbari, Z., & Girija, S. (2023). A systematic survey on multimodal emotion recognition using learning

algorithms. *Intelligent Systems with Applications*, *17*, 200171. https://doi.org/10.1016/j.iswa.2022.200171

[16] Canedo, D., & Neves, A. J. (2019). Facial expression recognition using computer vision: A systematic review. *Applied Sciences*, *9*(21), 4678. https://doi.org/10.3390/app9214678

[17] Khan, F. (2018). Facial expression recognition using facial landmark detection and feature extraction via neural networks. *arXiv Preprint:1812.04510*. https://doi.org/10.48550/arXiv.1812.04510

[18] Zhou, N., Liang, R., & Shi, W. (2021). A lightweight convolutional neural network for real-time facial expression detection. *IEEE Access*, *9*, 5573–5584. https://doi.org/10.1109/ACCESS.2020.3046715

[19] Duncan, D., Shine, G., & English, C. (2016). Facial emotion recognition in real time. *Computer Science*, 1–7.

[20] Ghimire, D., & Lee, J. (2014). Extreme learning machine ensemble using bagging for facial expression recognition. *Journal of Information Processing Systems*, *10*(3), 443–458. https://doi.org/10.3745/JIPS.02.0004

[21] Novotny, M., Tykalova, T., Ruzickova, H., Ruzicka, E., Dusek, P., & Rusz, J. (2022). Automated video-based assessment of facial bradykinesia in de-novo Parkinson's disease. *npj Digital Medicine*, *5*(1), 98. https://doi.org/10.1038/s41746-022-00642-5

[22] Asgarian, A., Zhao, S., Ashraf, A. B., Browne, M. E., Prkachin, K. M., Mihailidis, A., ..., & Taati, B. (2019). Limitations and biases in facial landmark detection–An empirical study on older adults with dementia. In *CVPR Workshops*, 28–36.

[23] Wu, X., Hong, D., & Chanussot, J. (2023). UIU-Net: U-Net in U-Net for infrared small object detection. *IEEE Transactions on Image Processing*, *32*, 364–376. https://doi.org/10.1109/TIP.2022.3228497

[24] Krishnan, A., Almadan, A., & Rattani, A. (2020). Understanding fairness of gender classification algorithms across gender-race groups. In *19th IEEE International Conference on Machine Learning and Applications*, 1028–1035. https://doi.org/10.1109/ICMLA51294.2020.00167

[25] Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z., & Matthews, I. (2010). The extended Cohn-Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 94–101. https://doi.org/10.1109/CVPRW.2010.5543262

[26] Lundqvist, D., Flykt, A., & Öhman, A. (1998). *Karolinska directed emotional faces* [Database record]. APA PsycTests. https://doi.org/10.1037/t27732-000

[27] Goodfellow, I. J., Erhan, D., Carrier, P. L., Courville, A., Mirza, M., Hamner, B., ..., & Bengio, Y. (2013). Challenges in representation learning: A report on three machine learning contests. In *Neural Information Processing: 20th International Conference*, 117–124. https://doi.org/10.1007/978-3-642-42051-1_16

[28] Permanasari, Y., Ruchjana, B. N., Hadi, S., & Rejito, J. (2022). Innovative region convolutional neural network algorithm for object identification. *Journal of Open Innovation: Technology, Market, and Complexity*, *8*(4), 182. https://doi.org/10.3390/joitmc8040182

[29] Hu, X., Jeon, Y., & Gwak, J. (2023). Heterogeneous ensemble approaches for robust face mask detection in crowd scenes. *Journal of Computational and Cognitive Engineering*, *2*(4), 343–351. https://doi.org/10.47852/bonviewJCCE3202478

[30] Guo, X., Zhang, Y., Lu, S., & Lu, Z. (2024). Facial expression recognition: A review. *Multimedia Tools and Applications*, *83*(8), 23689–23735. https://doi.org/10.1007/s11042-023-15982-x

[31] Loukadakis, M., Cano, J., & O'Boyle, M. (2018). Accelerating deep neural networks on low power heterogeneous architectures. In *11th International Workshop on Programmability and Architectures for Heterogeneous Multicores*, 1–14.

How to Cite: Krishnasamy, N., Zade, N., Khambholia, D., Henry, R., & Gupte, A. (2025). Ensemble Deep Learning Framework for Hybrid Facial Datasets Using Landmark Detection: State-of-the-Art Tools. *Journal of Computational and Cognitive Engineering*, *4*(4), 535–545. https://doi.org/10.47852/bonviewJCCE52024451