IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Traitement du Signal
Vol. 42, No. 1, February, 2025, pp. 519-530
Journal homepage: http://iieta.org/journals/ts

Fusion of Ensemble Technique with CNN Model to Equilibrate Prediction of Face Emotions in Real-Time

Dipti Pandit[1,2*], Sangeeta Jadhav[3]

[1] Electronics and Telecommunication Department, Vishwakarma Institute of Information Technology, Pune 411048, India
[2] Electronics and Telecommunication Department, D Y Patil College of Engineering, Pune 411044, India
[3] Information Technology, Army Institute of Technology, Pune 411015, India

Corresponding Author Email: dipti.pandit@viit.ac.in

Copyright: ©2025 The authors. This article is published by IIETA and is licensed under the CC BY 4.0 license (http://creativecommons.org/licenses/by/4.0/).

https://doi.org/10.18280/ts.420144

Received: 6 June 2024
Revised: 28 August 2024
Accepted: 6 September 2024
Available online: 28 February 2025

Keywords:
ensemble techniques, voting normalization, face emotions prediction, data augmentation, deep CNN models

ABSTRACT

Researchers are exploring the use of humans' remarkable skill in identifying and distinguishing emotions for computerization purposes. Although face emotion prediction has extensive practical applications, it remains a challenging field of study due to its dependency on subjective factors. Despite age and occlusions, the method for equilibrate prediction of all fundamental facial emotions is presented in this study. A methodology for real-time facial emotion prediction utilizing an ensemble classifier, incorporating deep CNN models as primary base classifiers, while tackling the issue of imbalanced datasets., the CK+ and JAFFE datasets are synthetically enhanced through image expansion approaches. A metaclassifier utilizing a combination of majority and relative voting techniques is employed at level 2 to improve the precision of individual emotions. The proposed method is tested using the internet's randomly selected facial expression images, demonstrating enhanced overall accuracy. Furthermore, cross-validation on the FER2013 dataset is performed utilizing the proposed ensemble fusion method.

1. INTRODUCTION

Human communication can be textual, verbal, nonverbal, through music, or any blend of these. In each mode, efforts are now being made to automate expression recognition. Facial expressions contribute 55 percent of human communication, with vocal communication contributing 38 percent and spoken communication contributing only 7 percent. Human expressions are the signals that are used to transmit a message, and emotions are the messages that are conveyed through expressions. Machine learning and computer vision have an impact on how expressions from text, multimodal sources, and music are encoded. Speech, facial expressions, gestures, and head-eye movements are all used to discern emotions. Researchers have identified 27 key expressions for expressing our emotions, seven of which are basic emotions. A variety of contexts, spanning biology, brain science, internet sites and commercial testing, ethnology, surveillance, psychology, and plenty of others, assessing audience feedback during seminars, lectures, and interrogations, intercept emotional faces and interpret them. Facial expression identification can be automated using muscle movement analysis, dimensional, holistic, vocal, and other approaches. Erroneous computation, impulsive, emotional behavior, illumination variation due to head activities, registration methods potentially causing false registration, occlusions caused by accessories such as googles, specs, and even camera arrangements, identity errors due to the independent emotional intensity of subjects, and wrinkles caused by aging are the main challenges in automating emotion or expression prediction. Despite decades of research, many concerns related to human-computer interaction (HCI) remain unresolved, particularly concerning determining which indicators and expressions should be examined for message encoding. This is due to the persistent challenge of extrapolating classifiers to undiscovered individuals with varying actions and facial characteristics such as wrinkles induced by aging, brows that can slip, or incorrect emotions.

The Viola-Jones approach, which builds on the Harr technique, can be used to identify and track faces. The methods, like feature extraction according to the data encoded, Appearance-based, and Shape-based, found much work done. Different pre-processing strategies are investigated depending on that feature selection technique to reduce undesirable deformities and for adjustments required owing to artifacts and occlusion. Numerous times, Gabor magnitude is utilized as a feature that is unaffected by misalignment [1, 2]. LBP, which was initially developed for texture analysis, is extensively used as a face analysis tool. LBP is distinguished by its tolerance to strobing lights, supercomputing convenience, and sensitivity to local regions while remaining reliable to volatility in face alignment [3]. However, some of the patterns have been demonstrated to be susceptible to encoding noise or false edifices and are inherently resistant to rotations. LBP is modified in a variety of ways, including LBP-TOP, LGBP, and LGBP-TOP [4, 5]. Based on the information gain rate and the methods of threshold selection and random dropout, a multi-scale and multi-region vector triangular texture feature extraction approach is employed to optimize the feature space

[6]. Machine learning algorithms, in general, address three problems: regression, grouping, and classification. Generally, based on the types and categories accessible, one strategy is chosen from among those available. For classifying and predicting facial emotions, neural networks, and other machine learning approaches like Decision trees, KNN [7], SVM [8], Random-forest, and deep learning models [9-11] are developed. Many algorithms are combined to improve performance to generate a hybrid optimization algorithm [12]. Due to their capacity to extract pertinent abstractions from data, deep learning-based algorithms, especially those based on CNNs, have recently demonstrated tremendous success in image-related tasks.

The crucial factor that enables deep learning is the availability of powerful computing systems and vast quantities of data that can be used to construct large neural networks. A DCNN employs high-dimensional images and numerous hidden convolutional layers. Every model has connections and layer configurations that are vastly diverse, which makes input and training extremely challenging. The architecture model VGG-19 [9], was created with 19 convolution layers, each with a higher accuracy level. Then, 8 learned layers, including 5 convolutional layers and 3 fully connected layers, were trained for the AlexNet model using the ImageNet dataset. The DCNN [13] network was then used to recognize facial emotion, which was trained using EmotiW 2015. According to reference [14], for FER, most scholars choose CNN-style models due to various factors, including the need for a large database, superior resolution images, generalizing the model, and the difficulty in increasing the accuracy, ultimately leading to the vanishing gradient problem. Emotion classification has always been a problem, and it's dealt with in various traditional and inventive methods. It's simple for complicated models like CNNs to overfit the data when working with limited datasets for image-based static face emotion recognition. A transfer learning technique is used to have a wide-capacity classifier and a predictor on tiny datasets, whereby the weights for a CNN are combined with those generated by a network nurtured for a specific job before fine-tuning parameters using the target dataset. However, since mention of all the techniques used till now is difficult to fit here, the authors suggest the best comprehensive papers related to face emotion recognition [15].

Identifying an individual's emotion can be tough for humans due to minute variances in feelings among the more complicated emotions. As an outcome, a classifier must have efficient features that have been fine-tuned and optimized for this specific purpose to produce effective predictions. A potential solution is to have enough training data for distinct classifiers, which isn't always possible or practical, and to take a holistic approach to facial emotion registration. We have tried to touch on the holistic approach with hybrid techniques using ensemble methods on CNN algorithms to equilibrate the face emotion predictions for all the basic emotions. Ensemble methods combine the predictions of multiple base estimators that are generated using a particular learning approach. This is done to enhance generalizability or resilience compared to a single estimator. In reference [16], the ensemble technique is used for the classification of power quality, weather forecasting in reference [17], and topic categorization in references [18, 19]. These worked on base classifiers like Multi-Layer perceptron, Logistic regression, KNN, Naïve Bayes, and SVM, followed by the metaclassifier. Our work's most major contribution can be summed up as follows:

1) Perform different experiments using diverse deep learning architectures;

2) Predict the 7 emotions with only facial components using the ensemble method with most of the deep learning models as base classifiers regardless of dataset size, illumination, face alignment, age, and occlusion of the subjects in the dataset;

3) Comparing the performance with other techniques for validation of the proposed method with random internet images as well as on the newly created dataset.

The research focuses on developing a robust and balanced real-time face emotion prediction approach using an ensemble technique with deep CNN models. The innovations include leveraging image augmentation to address dataset imbalances and employing a metaclassifier to enhance accuracy through majority and relative voting strategies. Compared to existing methods, the study also demonstrates improved prediction accuracy on datasets like CK+, JAFFE, and FER2013, and validates the approach using random internet images, showcasing its real-world applicability.

The following is how the paper is organized: Section 2 includes the proposed work overview, the flow from data preparation to deep learning networks, and the ensemble technique used as a meta-classifier. In Section 3, the experiments and results of the proposed work are discussed concerning and evaluated with related work. Section 4 concludes the paper with an observation and probable impending research orders.

2. PROPOSED WORK

The predictions of numerous methods are integrated into the ensemble method using learning procedures to increase generalizability or resilience compared to a single estimator. Ensemble learning can be categorized into three algorithms based on distinct integration strategies: averaging, boosting, and stacking. The essential principle of the averaging approach is that multiple estimators are created independently, and then their estimates are averaged. Because its variance is decreased, the combined estimator outperforms any solo estimator. Boosting also considers homogeneous weak learners by successively training them in a highly adap% manner and combining them using specialized deterministic algorithms. Stacking, disparate bagging, and boosting consider assorted feeble learners and use a meta-learner to combine different classification models. Recently, the ensemble method was observed by Vandana and Marriwala [20], where the different CNN algorithms for face emotion recognition are used with an accuracy level of 75.2%. The proposed work ensemble learning and prediction framework comprises deeper algorithms that are tested and trained on two completely different datasets, with 3 levels as levels 0, 1, and 2 as shown in Figure 1.

The most important and primary stage of the entire process is database preparation. The proposed technique was evaluated using two widely executed facial emotion databases, Japanese Female Facial Expressions and Extended Cohn-Kanade. JAFFE and CK+ exhibit distinct distributions for the various emotions, with JAFFE having an even distribution and CK+ having an uneven distribution for all of the various emotions. The CK+ database contains subjects aged 18 to 50 years, with 69% being female. 81% of the participants in the database are European Americans, 18% are African Americans, and 6% belong to other groups [21]. The grey scale dimensions with

8-bit or 24-bit color values are arrayed in 640 by 480-pixel arrays of converted images from baseline to target. The CK dataset is now referred to as the CK+ dataset because it has various areas of changed intensity that are of interest, such as frame-after-frame action component intensities data annotation and an overall of fourteen action units with impulsive footage and original participants, as well as annotations and tags for non-basic emotions. JAFFE comprises seven expressions posed by Japanese ladies, six of which are fundamental and one of which is neutral [22].

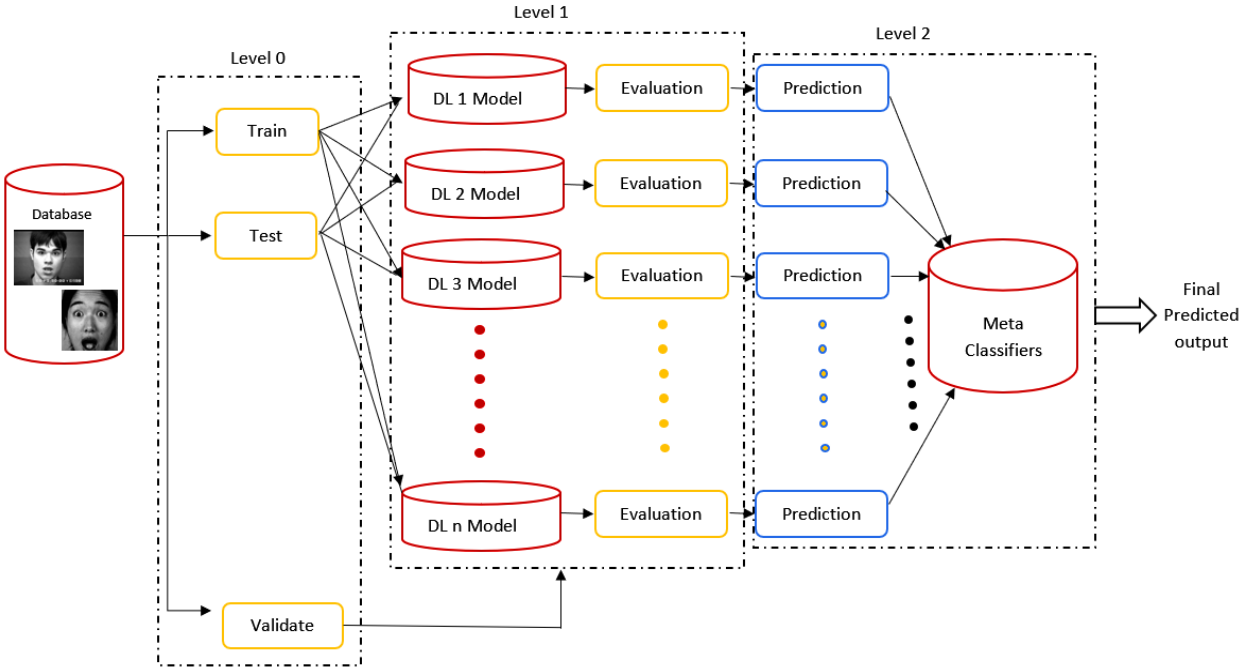

Figure 1. Proposed model structure for face emotion recognition

The methodology section employs ensemble techniques with CNN models, specifically integrating deep learning architectures such as DenseNet, ResNet50V2, InceptionResNetv2, VGG19, and MobileNet. The models were implemented with detailed pre-processing steps, including dataset augmentation using rotation, shifting, and flipping to address imbalanced data. Each model was fine-tuned with specific hyperparameters, such as a batch size of 256, learning rates initialized at 0.001 and adjusted using Adamax optimizer. The rationale for selecting these architectures was based on their proven performance in feature extraction and generalization across varied datasets. By combining these models through majority and relative voting in an ensemble framework, the study aims to enhance prediction accuracy, particularly for real-time emotion detection. Including these details ensures clarity and supports reproducibility.

Numerous hyperparameters and parameters have to be chosen when constructing a model using deep learning for specific tasks to ensure that the system can complete the task. However, when dealing with a small dataset for an inert facial expression recognition task that contains diverse data, including modeled and non-modeled types, multifaceted algorithms, like CNNs, can readily overfit the data.

Figure 2. Generation of the dataset using rotation and flipping technique on CK+ and JAFFE dataset

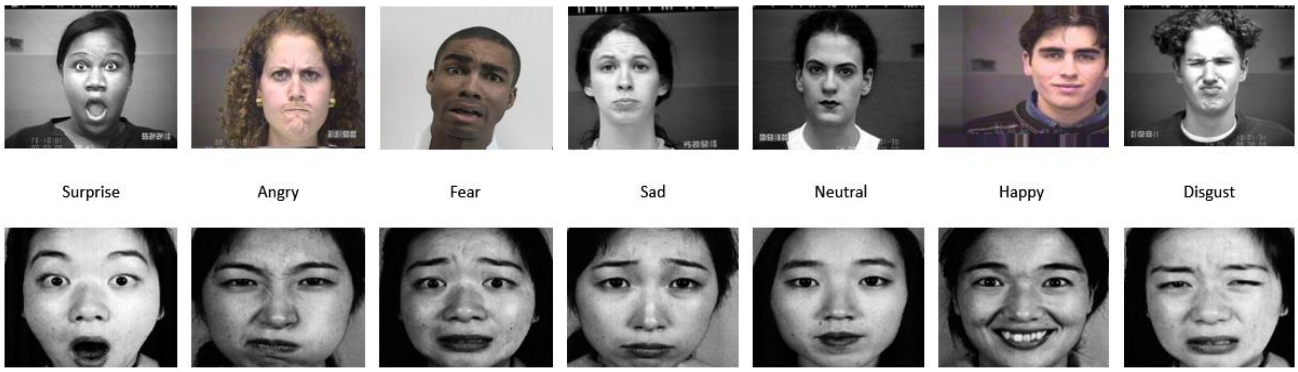

Figure 3. Sample images from the CK+ and JAFFE datasets after augmentation

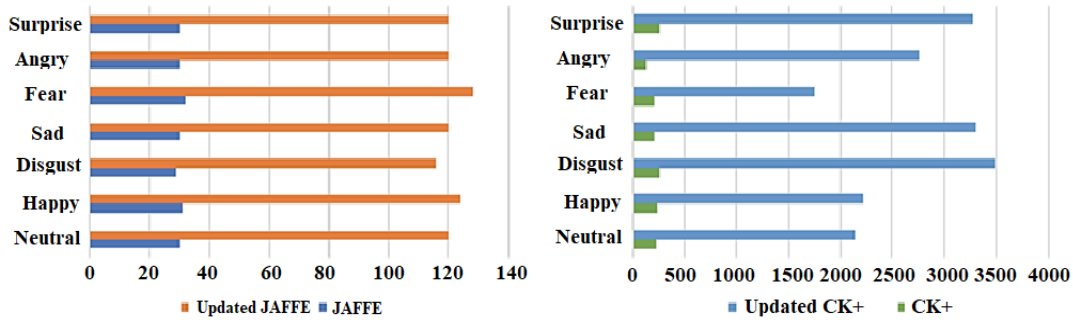

Figure 4. Dataset comparison of before and after augmentation for JAFFE and CK+ dataset

To expand the database size for enormous training, testing, and validation, new images are generated synthetically by rotation, shifting, and flipping actions from image augmentation, as demonstrated in Figures 2 and 3. Figure 4 shows that the total image in the JAFFE and CK+ databases has nearly increased by 6 to 9%.

Data augmentation methods, such as rotation, shifting, and flipping, were employed to enhance the CK+ and JAFFE datasets synthetically, addressing the issue of imbalanced data and limited sample size. These techniques expand the diversity of the training set, enabling the model to generalize better and reducing the risk of overfitting. By introducing variability in facial expressions and orientations, the model becomes more robust in recognizing emotions under different conditions, ultimately leading to improved accuracy and performance in real-time emotion prediction as shown in Figure 4.

After that, the image is cropped to eliminate the backdrop so that it is left with expression-specific elements. To get the features in distinct photos in the same place, a down-sampling process is used. This database is divided into three different phases of the system. The enhanced dataset with 7 different classes of emotions (6+1) is divided into training, testing, and validation sets at level 0. Here, the algorithm is trained and batches are created using the training dataset, whereas the test dataset is used to generalize the error and precision of the final algorithm, and the validation dataset is used to fine-tune the model as well as assess bias and variation.

2.1 Level 0

In level 0, the projected approach for facial emotion prediction, dataset splitting is done into training, validation, and testing sets of data since the CK+ and JAFFE datasets do not offer a quantified split. After creating training data consisting of grayscale photos of faces with their corresponding expression labels, the model learns a set of weights for the network. A few images are used to identify the final optimal set of weights from a collection of training completed with samples presented in varied order to ensure that the training performance is unaffected by the order in which the examples are presented. The validation dataset is utilized to assess the model's prediction error, where the validation score is computed. The validation needs are extensive, and the loss rate needs to be low. The algorithm is assessed using the dataset and the loss function. The loss function will have a high value if the prediction is low, and vice versa.

2.2 Level 1

While constructing a DNN for a specific objective, it is necessary to specify a number of configurations and parameters to make sure the network is appropriate for the job at hand. To address the challenge of developing a powerful classifier on a minor database, prior research in this field has focused on methods for multi-task learning. In the case of a CNN, this involves initialising the weights with those from a network that has been competent with associated tasks and then tweaking them using the intended dataset [9]. This technique exhausted manually training the learning network on the small dataset on a consistent basis, acquiring the concept known as "Knowledge Learning" [23, 24]. In level 1, as shown in Figure 5, a fine-tuning process is tested using a massive dataset termed D_A, which corresponds to task T_A.

Table 1. DNN Architectures comparison in terms of no of parameters and depth

Sr. No	Methodology	Depth	Parameters
1	VGG19	19	19.6 billion
2	Dense-Net	201	20 million
3	ResNet50V2	50	49 million
4	InceptionResNetv2	164	64 million
5	MobileNet	28	16 million

Firstly, a deep learning model is trained with D_A for T_A until good accuracy is achieved. The model is then used as a pre-trained model to fine-tune with multiple datasets in the next steps. We have tested the various networks and obtained a common inference. To begin, we build a CNN to train a deep model for T_A with D_A. The model performs well on the T_A and is then used as a pre-trained model in the following steps to fine-tune with numerous datasets. We tested the various networks and came up with a common conclusion. To describe the inferences, we further describe five models that are proposed for ensemble in the paper with different architectures publicized in Table 1. VGG19 [25], (19 layers) a pyramidal network with huge nethermost layers contiguous to the image and profound topmost layers, is characterized by its hierarchical arrangement. VGG has proven to be an excellent model for evaluating a specific job, however, due to the enormous number of factors, training takes an exceptionally long period (about 19.6 billion). ResNet50v2 [26], is one of the monster architectures that consists of multiple subsequent residual modules. Increasing the layers actually didn't improve accuracy and caused a vanishing gradient, which was solved using batch normalization. To improve accuracy, an identity connection was added between the layers. The pre-activation type of ResNet used in the proposed method has eliminated the remaining nonlinearity, paving the way from input to output in the manner of an identity connection. The total parameters are around 49 million, of which 27 million are

trainable. InceptionResNetv2 [25] is not a sequential architecture. In a single layer, multiple types of feature extractors are present, which helps in better performance. It is a CNN architecture that is built on the inception family but incorporates residual connections, i.e., replaces the filter concatenation stage of the inception architecture. This actually reduced the trainable parameters to only 13 million parameters out of 64 million in total. The Dense-Net [27] model includes a pristine connection that helps to alleviate the problem of gradient disappearing because each existing layer is linked to the prior one. Every network topology generates k feature maps, as revealed by the growth rate k. MobileNet, like InceptionResNet, reduces the trainable parameters to 2 million out of 16 million to achieve the training quickly. MobileNet [28], an im-ponderous DNN's reliance on depth-wise modular filters, which combine depth- and point-wise convolution filters, greatly eliminates computation. Point-wise convolution filters merge the result of the depth-wise convolution process linearly with 1 * 1 convolution, in contrast to depth-wise convolution filters, which conduct a single convolution on every input channel.

DNN techniques employ the fine-tuning technique, in which certain trained layers are frozen and a small number of layers are trained using a customized dataset. Using the training dataset for training all the different deep learning models, where the pre-trained models have trainable and nontrainable parameters. These trainable parameters are now used on the D_B dataset for a T_B task i.e., emotion recognition in our approach, which reduces development efforts. The k-fold cross-validation (CV) approach is commonly used to train respective pre-trained models, also called base classifiers, during the training phase to minimize the risk of over-fitting. These base classifiers' outputs are recorded in a hierarchical data format version 5 (HDF5). HDF5 files are metadata files containing a few groups of metadata along with the dataset, illustrated in Figure 6. HDF5 files consist of groups that are containers for datasets and other groups and can be used to organise data according to its logical structure. HDF5, an open, accessible file format that allows large, complicated, heterogeneous data that requires random and parallel access and is stored in a directory-like framework, is a file format that supports large, complicated, diverse data that requires random and parallel access. Using Figure 7, optimiser weights for the training database after level 1 in the HDF5 file format are demonstrated with the shape of the data according to the groups of m, v, and vhat. A powerful attribute of HDF5 is data slicing, by which a particular subset of a dataset can be extracted for processing, which helps in using very large datasets efficiently. The deep learning models, after training, generate huge trained parameters that can be further used for prediction. All these HDF5 files for each algorithm are saved individually. Every time, the training algorithms necessitate a massive amount of processing power.

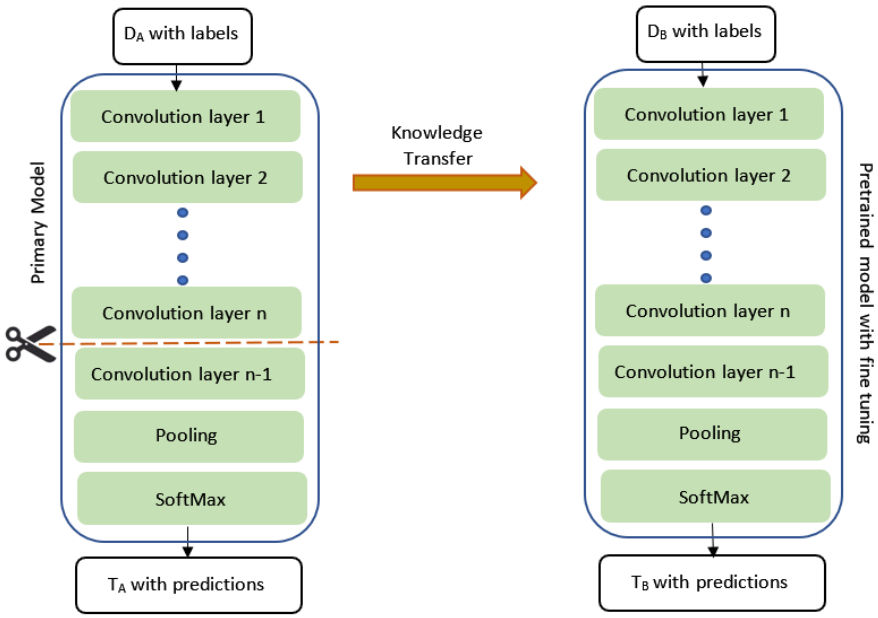

Figure 5. Knowledge transfer concept for TB with DB

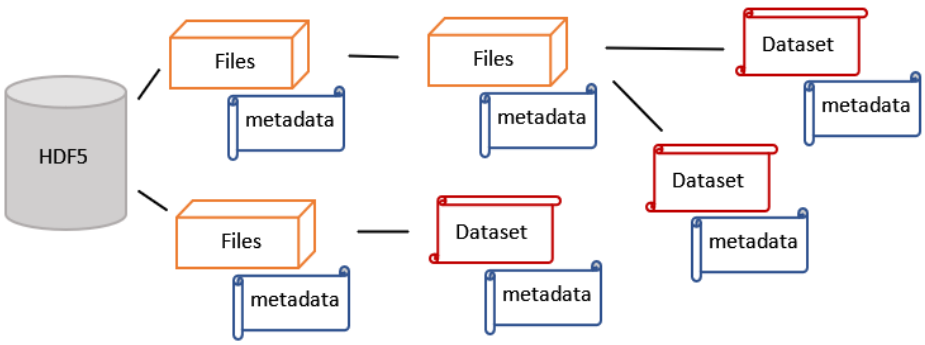

Figure 6. Structure of HDF5 files

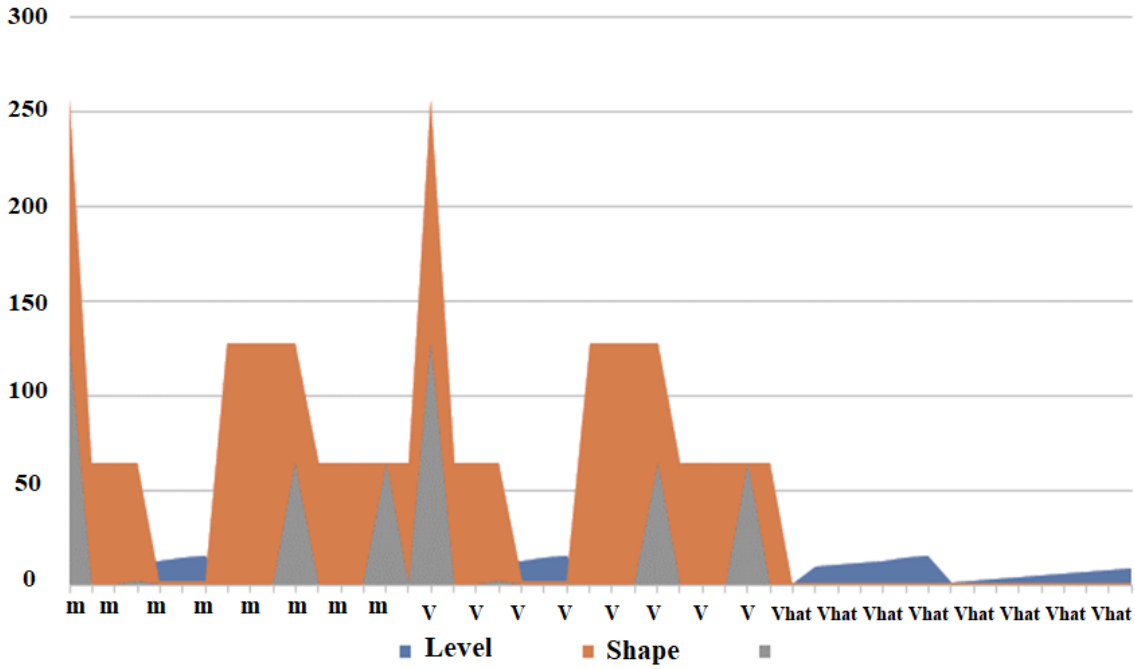

Figure 7. Optimiser weights for the training database after level 1 in the HDF5 file format, demonstrating the shape of the data according to the groups of m, v and vhat respectively

2.3 Level 2 (ensemble fusion technique)

In level 2, when adopting ensemble methods, the proposed methodology reduces the computation required by storing and reusing the trained model and increasing the overall recognition accuracy. The voting classifier predicts class labels by implementing a majority vote or the average projected probabilities. It combines conceptually distinct machine learning classifiers.

2.3.1 Majority voting

The ensemble's anticipated target label is the mode of the individual predicted labels' distribution, which is from all the individual classifiers, the most common prediction is considered the final output.

Let's consider, h as the classifier, and y as the output

$$\hat{y} = mode\{h_1(x), h_2(x), \dots\dots, h_n(x)\} \quad (1)$$

where, $h_i(x) = \widehat{y_i}$ and so the above equation can be written as,

$$\hat{y} = mode\{\widehat{y_1}, \widehat{y_2}, \dots\dots, \widehat{y_n}\} \quad (2)$$

By applying the probability mass function of a binomial distribution and the assumption that h classifiers predict the same class label, it is possible to determine the likelihood of making an inaccurate forecast using the ensemble technique.

We can say that,

$$P_h = \left(\frac{n!}{(n-h)h!}\right) \in^h (1-\in)^{n-h} \quad (3)$$

where, h > [n\2] and n is the number of classes. The total error prediction of an ensemble classifier will always be less than the base classifier error if the base error is always less than 50%. The ensemble method actually lowers the error more compared to using only a base classifier.

2.3.2 Relative voting

In relative voting, which is also called a weighted average ensemble, we use predictive probability instead of the class label, and we add a weighting component to the majority vote. The predictions are totaled and weighted according to the importance of the classifier, and the target label with the highest sum of the weighted probabilities gets the vote.

$$\hat{y} = argmin_j \sum_{i=1}^{n} w_i p_{i,j} \quad (4)$$

where, w_i is the weight parameter, while p_i. j is the ith classifier's projected class association probability for the class label j.

$$\widehat{y_f} = \frac{1}{n}\sum_{i=1}^{n} \widehat{y_i} \quad (5)$$

2.3.3 Ensemble fusion

Due to the unbalanced dataset for a few classes, there is a large probability that the performance of individual classes can vary. The main focus is to improve the prediction of weaker classes and to evaluate random real-time images for emotion prediction. To reduce the false positive and false negative, relative voting weights like [0.5, 1, 1, 1, 0.75] are allotted on try and error bases. The combination of majority (hard) voting and relative (weighted) voting attempts to make the work performance neutral for unbalanced dataset classes and balances out the weaknesses of all meta-classifiers:

$$\widehat{y_{final}} = \frac{1}{2}\left(\widehat{y_f} + \hat{y}\right) \quad (6)$$

To normalize the final predictions $\widehat{y_f}$, the standard mean is calculated to get the final predictions from the forecasts received from the voting ensemble methods. This final

prediction $\widehat{y_f}$, is then used as input feature weights for a metaclassifier as generalizer [29]. For generalization a learning set of m pairs: $\{x_k \in R^n, y_k \in R\}$, where, $1 \leq k \leq m$. Prediction is $x_k \in R^p$ or R, here, p=1, positive integer and one for better understanding. $\{g_i\}$, $1 \leq i < x$, where, g_i is complete set, x is the learning set input. If the generalizer returns the appropriate y_i, whenever w is equal to 1 of the x_i, in the learning set. Then we say the generalizer reproduces the learning set. This attempts to reduce covariance among base models while keeping the ensemble's modification and bias terms constant. Thus, a real-time image is given as input to the system, and all the respective algorithms will predict the emotions based on the confidence levels of both ensemble classifiers.

3. EXPERIMENTATION AND RESULTS

We conduct the classification experiments using the dataset for all seven emotions and later validate the complete algorithm for real-time images randomly picked from the internet. A fine-tuning methodology is tried in level 1, where the base classifier algorithms are trained for facial emotion prediction on more than 2000 picture sequences from over 200 people, aged between 18 and 30, comprising the CK+ and JAFFE datasets. The Cohn-Kanade collection includes, respectively, male and female facial expression image sequences for all of the six basic emotions. The last two photos from each sequence where the expression is at its maximal intensity were chosen in our experiments. The number of instances for each phrase fluctuates after the augmentation, depending on its availability. In our experiments on the CK+ database, we used 1522 images in total: neutral (231), happy (240), disgust (253), sad (211), fear (205), angry (126), and surprise (256). Whereas for the JAFFE database, we use 848 total images: neutral (120), happy (124), disgust (116), sad (120), fear (128), angry (120), and surprise (120). The image is resized with 224×224 input dimensions, and RGB images, followed by convolution and max-pooling layers. With all three fully linked layers and one SoftMax layer, the Relu activation function is applied. Batch sizes for all the algorithms are set to 256 and 20 with 50 and 15 epochs for the CK+ and JAFFE datasets, respectively.

The Categorical cross-entropy loss and Adamax loss optimizers are used during the training of the algorithms. The Adamax loss is computed based on the infinity norm, and when the error reaches a plateau, the default learning rate is divided by 10 from 0.001. The graphical depiction of training and validation losses used to diagnose the model's performance and help identify the need for tuning if necessary is shown in Figures 8-10.

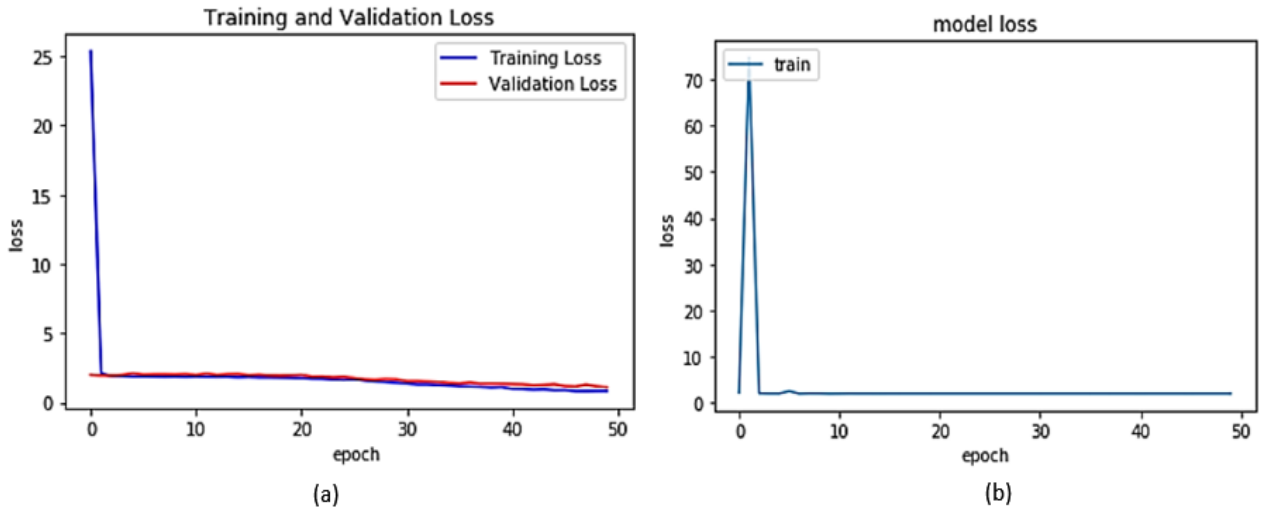

Figure 8. Training and validation loss for VGG19 Model using (a) CK+ and model loss for JAFFE (b)

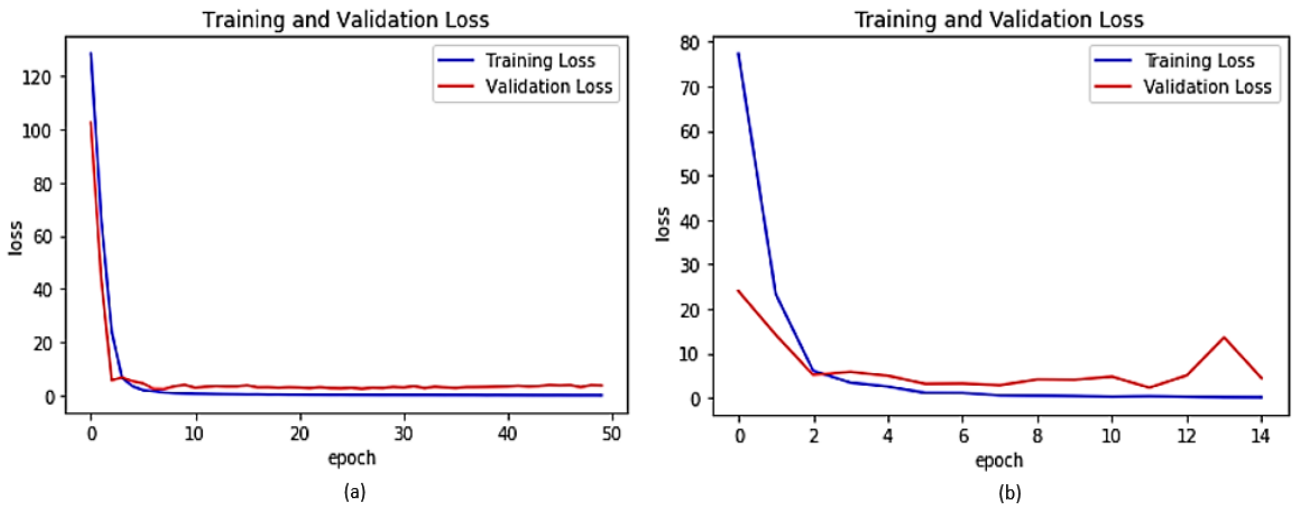

Figure 9. DenseNet201 Model using CK+ (a) and JAFFE (b) training and validation loss

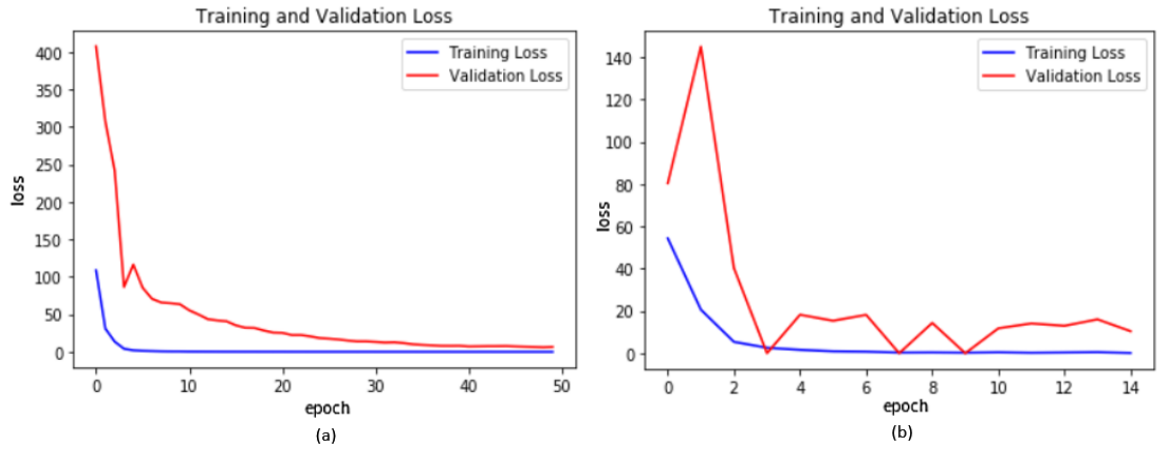

Figure 10. Training and validation loss using (a) CK+ and (b) JAFFE datasets for the MobileNet model

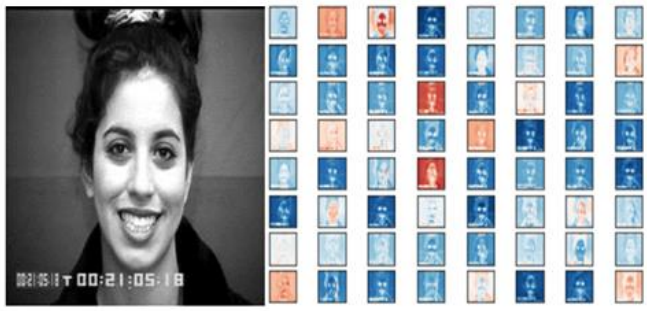

Figure 11. Visualization of layer 17 using DenseNet model

The evaluation metrics selected—accuracy, precision, recall, F1 score, and AUC ROC—were chosen for their comprehensive ability to assess different aspects of the model's performance, particularly in handling imbalanced datasets. Accuracy provides a general performance overview, while precision and recall offer insights into the model's ability to identify emotions, minimizing false positives and negatives correctly. The F1 score balances precision and recall, highlighting the model's robustness. AUC ROC assesses the classifier's discriminative ability across all classes, validating the effectiveness of the ensemble approach in predicting diverse emotions accurately.

Table 2. Performance of base classifier algorithms are evaluated in terms of accuracy using datasets CK+ and JAFFE

Base Classifier Algorithm	Dataset	Accuracy	AUC ROC	Precision
VGG19	JAFFE	13.33	50.18	17.11
	CK+	68.4	78.1	75.64
Dense-Net	JAFFE	97.45	85.91	67.28
	CK+	96.64	92.99	84.46
ResNet50V2	JAFFE	92.86	89.27	72.61
	CK+	99.37	96.61	90.13
InceptionResNetv2	JAFFE	95.63	90.49	75.32
	CK+	98.52	93.9	87.44
MobileNet	JAFFE	96.25	87.03	71.67
	CK+	99.21	94.72	87.23

The loss curve exhibits well-fitting learning curves for all the models by analyzing the generalization gap. The learning loss for VGG19, DenseNet, and MobileNet was 0.8354, 0.1521, and 0.0533, respectively. The lowest learning loss was observed for ResNet50V2 of 0.0365 for the CK+ dataset and DenseNet of 0.084 for the JAFFE dataset. The features of the DenseNet model for layer 17 are visualized using several layers in Figure 11. In Table 2, the performance of the base classifier algorithm performance is discussed with the AUC, ROC, accuracy, and precision parameters. The DenseNet and InceptionResNet models show the most significant results related to accuracy on both datasets. On the OpenCV platform, the outcome of all classifiers is assessed using both sets of data on 4x NVIDIA Professional Series Quadro P6000 RTX PCIe 3.0-24 GB and a Core i5 9th generation 16GB memory device. The CPU training time for each of the individual algorithms is illustrated in Figure 12 with respect to their individual accuracy as well as the time taken for training on the CPU and GPU. The CPU and GPU training periods for VGG19 are 22 hours and 6 hours, respectively. Dense-Net and InceptionResNetv2 CPU training takes 26 and 27 hours, respectively, while GPU training takes 7.5 and 8 hours. The Dense-Net, InceptionResNetV2 model has an average accuracy on two separate datasets of 97.05%, compared to VGG19's accuracy of only 40.85%.

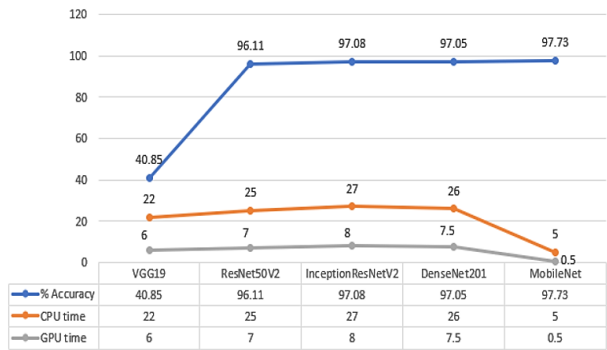

Figure 12. Overall learning time in hours and performance evaluation (accuracy) of base classifier Algorithms on CPU and GPU

The proposed approach's performance is compared to the performance of all other feasible approaches from the survey, as shown in Table 3. Here, the major goal is to suggest a method that is independent of dataset size.

The recognition rate for Boosted LBP and SVM

combination is 91.4% and 81% for the CK+ and JAFFE datasets, respectively, whereas the LDP and SVM classifiers show 93.4% and 85.4% recognition accuracy, respectively. NSLBP with multiclass Adaboost and Improved DTP with SVM are tried on only one dataset, which is CK+ and JAFFE, respectively.

Table 3. Comparison with our approach (Ensemble method using deep learning algorithm), with others for emotion prediction

Sr. No.	Methodology	CK+ Recognition Rate %	JAFFE Recognition Rate %
1	NSLBP + multiclass Adaboost [2]	97.7	-
2	Boosted LBP + SVM [30]	91.4 ± 3.8	81.0
3	LDP + SVM [31]	93.4 ± 1.5	85.4 ± 4.0
4	Improved DTP + SVM [32]	-	87.77 ± 7.15
5	FP +SAE [10]	91.11	90.47
6	Dense-Net	96.62 ± 0.2	97.45 ± 0.12
7	ResNet50V2	99.33 ± 0.4	92.86 ± 0.33
8	InceptionResNetv2	98.52 ± 0.1	95.63 ± 0.21
9	MobileNet	99.21 ± 0.2	96.25 ± 0.3
10	Ensemble method (our proposed method)	99.87 ± 0.1	98.82 ± 0.1

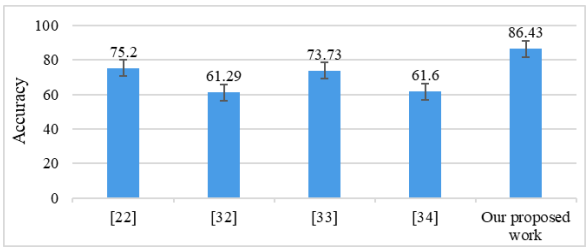

Figure 13. Result comparison with other ensemble methods used for face emotion recognition

The effectiveness of individual deep learning models is evaluated on both datasets. The proposed ensemble fusion methodology is cross-validated on the FER2013 dataset to compare its performance with other techniques. The FER2013 dataset contains approximately 30,000 images of various facial expressions, each with a compressed size of 48×48. Figure 13 summarizes the results of different methodologies applied in combination with the proposed ensemble method.

On the FER2013 dataset, reference [20] evaluates facial emotion recognition using the VGG, Inception, and ResNet algorithms, along with the discriminative DCN and AMN approach from the study by Kim et al. [33] and the deep CNN method from the study by Kim et al. [34]. In the study by Yu and Zhang [35], the SFEW dataset is used with a voting approach based on an image-based CNN architecture. It is also noted that algorithms such as SVM, decision trees, and random forests are frequently employed with stacking-based ensemble techniques.

The proposed approach is further evaluated using random real-time images for emotion prediction. Accuracy is calculated by dividing the total number of correct predictions by the total number of samples. While accuracy performs well on balanced datasets, it can be misleading for unbalanced data. To assess the quality of the classification, precision, recall, accuracy, and the F1 score are calculated. Precision, in particular, indicates the proportion of correctly predicted positive cases out of all predicted positive cases.

$$Precision = \frac{TP}{TP + FP} \tag{7}$$

Recall tells us how many of the actual positives we were able to predict correctly with our model.

$$recall = \frac{TP}{TP + FN} \tag{8}$$

F1 score is also known as the harmonic mean of precision and recall, as it captures both trends in a single value.

$$F1\ score = 2 \times \frac{precision \times recall}{precision + recall} \tag{9}$$

The proposed ensemble system exhibits a balanced F1 score of 98.62% with 98.83% recall and 98.45% precision on the CK+ dataset, as shown in Table 4. The proposed system's performance on the JAFFE database was also noticeably improved, with 88.96 percent precision and 84.23 percent recall, resulting in an average F1 score of 86.53 percent, as illustrated in Table 5. Happy, surprised, neutral, and disgusted expressions had the highest recognition rate, as seen in the table below. Despite having extremely identical facial motions, the sad and terror performances have also been warmly received.

Table 4. Comparison with Ensemble method using deep learning algorithm with individual models for face emotion prediction on CK+ dataset

	F1 Score	Recall	Precision
VGG19	61.73	52.14	75.64
Dense-Net	81.6	78.93	84.46
ResNet50V2	89.91	89.68	90.13
InceptionResNetv2	87.49	87.55	87.44
MobileNet	86.09	84.98	87.23
Ensemble method (our proposed method)	98.62	98.83	98.45

Table 5. Comparison with Ensemble method using deep learning algorithm with individual models for face emotion prediction on JAFFE dataset

	F1 Score	Recall	Precision
VGG19	14.57	12.69	17.11
Dense-Net	60.45	54.88	67.28
ResNet50V2	71.04	69.54	72.61
InceptionResNetv2	74.38	73.47	75.32
MobileNet	73.1	70.96	71.67
Ensemble method (our proposed method)	86.53	84.23	88.96

The error matrix, also known as the confusion matrix, describes how well the suggested workflow predicts facial emotion. The genuine positive values for the CK+ and JAFFE databases are high for all the emotions, as seen in Tables 6 and 7. Using the ensemble method for final prediction helped in processing time as well as the prediction accuracy for all the classes, especially minor classes like disgust, fear, and anger. To verify the technique, we gathered 57 random face images

from the internet with the relevant emotions: neutral, pleased, disgusted, sad, fearful, angry, and surprised, all of which were 10, 10, 6, 8, 6, and 10 in each case, as shown in Figure 14. The images depict people of all ages.

The prediction of emotion was validated with human labeling. The time duration for prediction of emotion was 0.578125 sec on the device with the configuration of a Core i5 9th generation with 16GB memory and 117KB of actual executable file. With an overall accuracy of 97.54 percent in predicting genuine emotion, Table 8 displays the confusion matrices created using the proposed ensemble approach on randomly selected images from the internet.

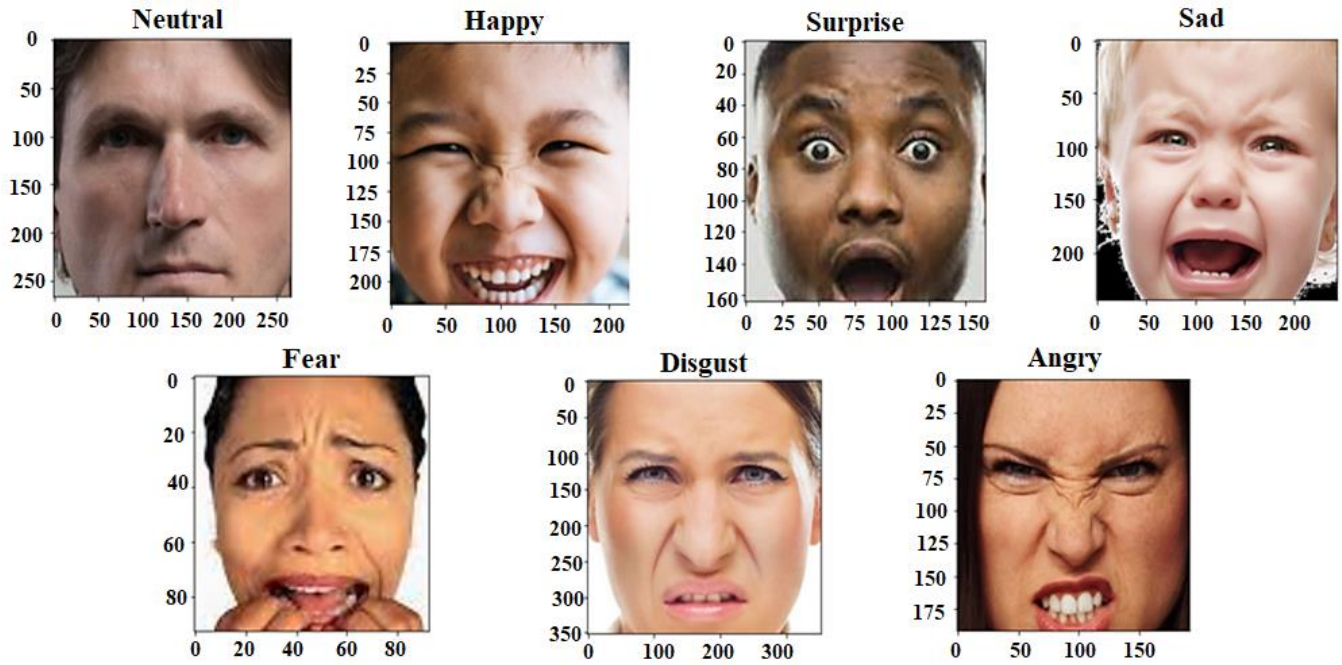

Figure 14. Example images from random internet for validation of the ensemble method for all seven emotions

Table 6. Confusion matrices on Extended Cohn-Kanade using the proposed ensemble approach

	Neutral	Happy	Disgust	Sad	Fear	Angry	Surprise
Neutral	99.95	0	0	0	0	0	0
Happy	0	99.72	0	0	0	0	0.28
Disgust	0	0.28	99.92	0	0.16	0	0
Sad	0	0	0	99.74	0.32	0	0
Fear	0	0	0.08	0	99.52	0.2	0
Angry	0.05	0	0	0.26	0	99.8	0
Surprise	0	0	0	0	0	0	99.72

Table 7. Confusion matrices on Japanese Female Facial Expressions using the proposed ensemble approach

	Neutral	Happy	Disgust	Sad	Fear	Angry	Surprise
Neutral	99.65	0	0	1.02	0	0	0
Happy	0	99.72	0	0	0	0	0.54
Disgust	0	0.28	97.88	0	0.54	0.7	0
Sad	0.35	0	1.45	97.86	0	0	0
Fear	0	0	0.52	0.16	98.92	1.05	0
Angry	0	0	0.15	0.96	0	98.25	0
Surprise	0	0	0	0	0.54	0	99.46

Table 8. Confusion matrices on random relevant images from the internet using the proposed ensemble approach

	Neutral	Happy	Disgust	Sad	Fear	Angry	Surprise
Neutral	97.52	0	0.56	1.32	0.47	0.13	0
Happy	0	98.86	0.31	0	0	0.83	0
Disgust	0.47	0.28	96.29	0.18	0.96	0.87	0.95
Sad	1.28	0	1.14	97.58	0	0	0
Fear	0.21	0	1.02	0.16	97.32	1.29	0
Angry	0.52	0.18	0.68	0.76	0.26	96.88	0.72
Surprise	0	0.68	0	0	0.99	0	98.33

Table 9. Paired t-test results for face emotion recognition methods

Dataset	t-Statistic	p-Value
CK+	-3.036	0.016
JAFFE	-3.514	0.008

The paired t-test results in Table 9 indicate statistically significant differences between the proposed ensemble method and other methods for the CK+ and JAFFE datasets, with p-values less than 0.05. This suggests that the proposed method outperforms the others in terms of recognition rates, validating its effectiveness in face emotion recognition.

4. CONCLUSION AND FUTURE SCOPE

This work presents how to predict facial emotions in real time using an ensembled classifier and deep learning techniques. The proposed system comprises deep CNN algorithms like DenseNet, ResNet, InceptionResNet, VGG19, and MobileNet as base classifiers to classify 7 human face emotions. Further, the performance of the proposed work is evaluated using real-time random images from the internet. The proposed normalization ensemble technique outperforms individual classifiers in terms of prediction performance, with F1 scores, Accuracy, Precision, and Recall values of 98.62 percent, 99.87 percent, 98.45 percent, and 98.83 percent, respectively, as well as neutralizing the prediction caused by imbalanced training data, especially for the minority class. To summarise, the suggested normalization ensemble classifier can be utilized to predict the face emotion classification in real time. Other than the 7 basic emotions, other emotions are yet to be worked on due to a lack of samples for training. But still, since the parameters required for prediction are optimized by using the pre-trained models followed by the ensemble technique, the overall size of the actually used model is reduced, making it simple to deploy on any device for real-time emotion prediction.

REFERENCES

[1] Cruz, A., Bhanu, B., Thakoor, N.S. (2013). Facial emotion recognition with anisotropic inhibited Gabor energy histograms. In 2013 IEEE international conference on image processing, Australia, pp. 4215-4219. https://doi.org/10.1109/ICIP.2013.6738868

[2] Rani, P.I., Muneeswaran, K. (2018). Emotion recognition based on facial components. Sādhanā, 43: 48. https://doi.org/10.1007/s12046-018-0801-6

[3] Saaidia, M., Zermi, N., Ramdani, M. (2016). Fuzzy linear projection on combined multi-feature characterisation vectors for facial expression recognition enhancement. International Journal of Signal and Imaging Systems Engineering, 9(4-5): 252-261. https://doi.org/10.1504/IJSISE.2016.078266

[4] Almaev, T.R., Valstar, M.F. (2013). Local gabor binary patterns from three orthogonal planes for automatic facial expression recognition. In 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction, Geneva, Switzerland, pp. 356-361. https://doi.org/10.1109/ACII.2013.65

[5] Cruz, A.C., Bhanu, B., Thakoor, N.S. (2014). Vision and attention theory based sampling for continuous facial emotion recognition. IEEE Transactions on Affective Computing, 5(4): 418-431. https://doi.org/10.1109/TAFFC.2014.2316151

[6] Tang, J.M., Mao, J.F., Sheng, W.G., Hu, Y.H., Gao, H. (2021). Texture feature extraction and optimization of facial expression based on weakly supervised clustering. Systems Science & Control Engineering, 9(1): 514-528. https://doi.org/10.1080/21642583.2021.1943725

[7] Ramirez, G.A., Fuentes, O., Crites Jr, S.L., Jimenez, M., Ordonez, J. (2014). Color analysis of facial skin: Detection of emotional state. In 2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops, Columbus, OH, USA, pp. 474-479. https://doi.org/10.1109/CVPRW.2014.76

[8] Brahnam, S., Chuang, C.F., Shih, F.Y., Slack, M.R. (2006). Machine recognition and representation of neonatal facial displays of acute pain. Artificial Intelligence in Medicine, 36(3): 211-222. https://doi.org/10.1016/j.artmed.2004.12.003

[9] Ng, H.W., Nguyen, V.D., Vonikakis, V., Winkler, S. (2015, November). Deep learning for emotion recognition on small datasets using transfer learning. In Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, Seattle Washingto, USA, pp. 443-449. https://doi.org/10.1145/2818346.2830593

[10] Zhao, X., Shi, X., Zhang, S. (2015). Facial expression recognition via deep learning. IETE Technical Review, 32(5): 347-355. https://doi.org/10.1080/02564602.2015.1017542

[11] Gudi, A., Tasli, H.E., Den Uyl, T.M., Maroulis, A. (2015). Deep learning based facs action unit occurrence and intensity estimation. In 2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), Ljubljana, Slovenia, pp. 1-5. https://doi.org/10.1109/FG.2015.7284873

[12] Al Amrani, Y., Lazaar, M., El Kadiri, K.E. (2018). Random forest and support vector machine based hybrid approach to sentiment analysis. Procedia Computer Science, 127: 511-520. https://doi.org/10.1016/j.procs.2018.01.150

[13] Sun, B., Li, L.D., Zhou, G.Y., Wu, X.W., He, J., Yu, L.Y., Li, D.X., Wei, Q.L. (2015). Combining multimodal features within a fusion network for emotion recognition in the wild. In Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, New York, USA, pp. 497-502. https://doi.org/10.1145/2818346.2830586

[14] Li, S., Deng, W. (2020). Deep facial expression recognition: A survey. IEEE Transactions on Affective Computing, 13(3): 1195-1215. https://doi.org/10.1109/TAFFC.2020.2981446

[15] Sariyanidi, E., Gunes, H., Cavallaro, A. (2014). Automatic analysis of facial affect: A survey of registration, representation, and recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 37(6): 1113-1133. https://doi.org/10.1109/TPAMI.2014.2366127

[16] Radhakrishnan, P., Ramaiyan, K., Vinayagam, A., Veerasamy, V. (2021). A stacking ensemble classification model for detection and classification of power quality disturbances in PV integrated power network. Measurement, 175: 109025.

https://doi.org/10.1016/j.measurement.2021.109025
[17] Venkatesan, R., Shirly, S., Selvarathi, M., Jebaseeli, T.J. (2023). Human emotion detection using DeepFace and artificial intelligence. Engineering Proceedings, 59(1): 37. https://doi.org/10.3390/engproc2023059037
[18] Campos, R., Canuto, S., Salles, T., de Sá, C.C., Gonçalves, M.A. (2017). Stacking bagged and boosted forests for effective automated classification. In Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, New York, USA, pp. 105-114. https://doi.org/10.1145/3077136.3080815
[19] Alali, A., Kubat, M. (2015). Prudent: A pruned and confident stacking approach for multi-label classification. IEEE Transactions on Knowledge and Data Engineering, 27(9): 2480-2493. https://doi.org/10.1109/TKDE.2015.2416731
[20] Vandana, Marriwala, N. (2022). Facial expression recognition using convolutional neural network. In Mobile Radio Communications and 5G Networks: Proceedings of Second MRCN 2021, pp. 605-617. https://doi.org/10.1007/978-981-16-7018-3_45
[21] 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 2010 ieee computer society conference on computer vision and pattern recognition-workshops, CA, USA, pp. 94-101. https://doi.org/10.1109/CVPRW.2010.5543262
[22] Shih, F.Y., Chuang, C.F., Wang, P.S. (2008). Performance comparisons of facial expression recognition in JAFFE database. International Journal of Pattern Recognition and Artificial Intelligence, 22(3): 445-459. https://doi.org/10.1142/S0218001408006284
[23] Atabansi, C.C., Chen, T., Cao, R., Xu, X. (2021). Transfer learning technique with VGG-16 for near-infrared facial expression recognition. Journal of Physics: Conference Series, 1873(1): 012033. https://doi.org/10.1088/1742-6596/1873/1/012033
[24] Nguyen, L.D., Lin, D., Lin, Z., Cao, J. (2018). Deep CNNs for microscopic image classification by exploiting transfer learning and feature concatenation. In 2018 IEEE International Symposium on Circuits and Systems (ISCAS), Florence, Italy, pp. 1-5. https://doi.org/10.1109/ISCAS.2018.8351550
[25] He, K., Zhang, X., Ren, S., Sun, J. (2016). Deep residual learning for image recognition. In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, pp. 770-778. https://doi.org/10.1109/CVPR.2016.90
[26] Ren, S., He, K., Girshick, R., Sun, J. (2016). Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(6): 1137-1149. https://doi.org/10.1109/TPAMI.2016.2577031
[27] Wang, W., Li, Y., Zou, T., Wang, X., You, J., Luo, Y. (2020). A novel image classification approach via dense-MobileNet models. Mobile Information Systems, 2020(1): 7602384. https://doi.org/10.1155/2020/7602384
[28] Sinha, D., El-Sharkawy, M. (2019). Thin mobilenet: An enhanced mobilenet architecture. In 2019 IEEE 10th annual ubiquitous computing, electronics & mobile communication conference (UEMCON), New York, USA, pp. 0280-0285. https://doi.org/10.1109/UEMCON47517.2019.8993089
[29] Wolpert, D.H. (1992). Stacked generalization. Neural Networks, 5(2): 241-259. https://doi.org/10.1016/S0893-6080(05)80023-1
[30] Shan, C., Gong, S., McOwan, P.W. (2009). Facial expression recognition based on local binary patterns: A comprehensive study. Image and Vision Computing, 27(6): 803-816. https://doi.org/10.1016/j.imavis.2008.08.005
[31] Jabid, T., Kabir, M.H., Chae, O. (2010). Robust facial expression recognition based on local directional pattern. ETRI Journal, 32(5): 784-794. https://doi.org/10.4218/etrij.10.1510.0132
[32] Tivatansakul, S., Ohkura, M., Puangpontip, S., Achalakul, T. (2014). Emotional healthcare system: Emotion detection by facial expressions using Japanese database. In 2014 6th Computer Science and Electronic Engineering Conference (CEEC), Colchester, UK, pp. 41-46. https://doi.org/10.1109/CEEC.2014.6958552
[33] Kim, B.K., Dong, S.Y., Roh, J., Kim, G., Lee, S.Y. (2016). Fusing aligned and non-aligned face information for automatic affect recognition in the wild: A deep learning approach. In 2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW): Las Vegas, NV, USA, pp. 1499-1508. https://doi.org/10.1109/CVPRW.2016.187
[34] Kim, B.K., Roh, J., Dong, S.Y., Lee, S.Y. (2016). Hierarchical committee of deep convolutional neural networks for robust facial expression recognition. Journal on Multimodal User Interfaces, 10: 173-189. https://doi.org/10.1007/s12193-015-0209-0
[35] Yu, Z., Zhang, C. (2015). Image based static facial expression recognition with multiple deep network learning. In Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, Seattle Washington, USA, pp. 435-442. https://doi.org/10.1145/2818346.2830595