Multimedia Tools and Applications (2023) 82:14689–14712
https://doi.org/10.1007/s11042-022-13940-7

An ensemble approach using a frequency-based and stacking classifiers for effective facial expression recognition

Rashmi Adyapady R.[1] · **B. Annappa**[1]

Received: 3 July 2021 / Revised: 15 July 2022 / Accepted: 12 September 2022 /
Published online: 4 October 2022

Abstract

Facial Expression Recognition is an essential aspect of human behavior to communicate effectively. A more profound understanding of human behavior, accurate analysis, and interpretation of the emotional content is essential. Hence, facial features play a crucial role as they contain beneficial information about facial expressions. A baseline architecture belonging to the EfficientNet family of models is explored for feature extraction. In this work, two novel strategies, the ensemble model using the frequency-based voting approach (FV-EffNet) and the stacking classifier (SC-EffNet), are proposed to enhance classification results' performance. The proposed system deals with both profile and frontal pose variations. The combination of deep learning models with a stacking classifier gave the best results of 98.35% and 98.06%, and the frequency-based approach used with the ensemble classifier achieved superior performance of 98.71% and 98.56% on Oulu-CASIA and RaFD datasets, respectively. The experiment results with the proposed methodology showed better performance than previous studies on Oulu-CASIA and RaFD datasets, making it more robust to pose variations.

Keywords Facial expression recognition · Emotion recognition · Deep learning · Machine learning · Ensemble model · Stacking classifier

1 Introduction

Facial expression conceded to be the best way to interact or communicate one's emotions and feelings. Charles Darwin [8] significant contribution was a consideration of discrete emotions; the second contribution was an emphasis on the face, as facial expression

✉ Rashmi Adyapady R.
rashmiadyapadyr.177co004@nitk.edu.in

B. Annappa
annappa@ieee.org

[1] Department of Computer Science and Engineering, National Institute of Technology Karnataka, Surathkal, India

contains the most valuable source of information; and the third contribution was revealing facial expressions of emotions as universal. The fourth observation was that emotions are not unique to humans but can be seen in other species. And his fifth contribution clarified why some movements correspond to a particular emotion. Thus, this began the evolution of the theory of Facial Expression Recognition (FER). Non-verbal components like facial expressions reveal 55% of the person's intention, verbal components, and vocal segments convey 7% and 38% of the communicated message respectively [11, 19]. Thus, this motivates the researchers to explore the area of FER efficiently. Every region of the face conveys some important affective information. Sometimes, it is challenging to separate the same subject's facial features in two different expressions, as they may share the same feature space [25]. There are issues with selecting appropriate features to distinguish individuals' emotions from various categories of emotions [59]. Expressions keep varying within the same culture [5, 35], and patterns may depend on environment settings, mood, and situations, making it difficult for machines to recognize them efficiently [25]. Variations in the face, facial occlusions, head poses, and illumination also degrades the overall system's performance. A generalized approach is needed that could overcome all these variations and help in building an efficient, robust system for recognition of expressions [54].

The composition of both feature extraction and classification techniques are essential for FER. The key challenges in efficiently recognizing facial expressions are a selection of efficient feature extraction and classification techniques. If features are sparse, the best classifier's performance would also gradually decrease [49]. The handcrafted features like Scale-Invariant Feature Transform (SIFT), Gabor, Local Binary Patterns (LBP), Histograms of Oriented Gradients (HOG) has achieved a breakthrough in various fields. These handcrafted low-level features work well on a small amount of training data and inadequate for extracting discriminative information. It is arduous to fine-tune these low-level features according to input data. These disadvantages of low-level features made it inefficient in recognizing facial expressions accurately in real-world applications. Thus, the deep learning models overcame these challenges and automatically learn from raw data, represent the data on multiple levels, and contain more abstract information. The rapid development in the deep learning field has impacted various areas, including FER, and has shown promising results in identifying expressions from facial images. Despite the success, computational cost remains high, imposing difficulty in availability and accessibility.

This work explores the EfficientNet [43] model, the high-quality model from the CNN group of models, to efficiently recognize facial expression from static images. It is efficient in terms of a lesser parameter (4M parameters) and achieves better accuracy than previous CNN models. The EfficientNet B0 baseline architecture is used as a feature extractor to improve the FER system's recognition rate and accuracy. Further, the features extracted from the EfficientNets intermediate layer are fed to machine learning classifiers for classification. A combination of deep learning models and machine learning classifiers effectively improves the ability of classification algorithms. Two ensemble models, EfficientNet B0, features fed to stacking classifier (SC-EffNet), and EfficientNet B0 features provided to machine learning classifiers based on the frequency of votes (FV-EffNet), are proposed to classify facial expressions into respective expression classes. Thus, the combination of multiple classifiers induces higher-level classifiers and tries to learn all possible patterns from the base classifiers [37], which further enhances the overall performance of FER.

The proposed work uses the EfficientNet model, which is computationally and memory-efficient compared to previous CNN models. The intermediate features of the models fed to machine learning classifiers using a frequency-based approach improved the system's accuracy even further. As a result, we chose the top five best weights and their corresponding intermediate features, which we fed into machine learning classifiers to test the system's performance. Additionally, to improve the model's performance, a stacking classifier (an ensemble approach) was used. The meta classifier analyzed the pattern of base classifiers and learned from their errors before making the final prediction. By integrating the outputs of base classifiers, the ensemble model makes accurate predictions, reduces over-fitting, reduces the risk of selecting a single classifier, and achieves good results. There is no work related to FER using best model weights, the extraction of EfficientNet features, and the stacking classifier for classification to the best of our knowledge.

In affective computing, emotion recognition from video data is the current issue [34]. Even though the amount of information obtained from the video signals is comparatively more. The quickness and variability in dynamics (rapid changes in the intensity of expressions from onset to peak and to offset state) of video sequences pose additional challenges, making it challenging to recognize the expressions in correlated frame sequences compared to static image analysis. Many recent works, including [34], have attested that FER on static images is still an active research area. This work focuses on FER based on static images rather than video sequences. The main research contributions of this study are summarized as follows:

1. Different from previous approaches, the top five best weights and their respective intermediate features are fed to the proposed ensemble models. Thus, the frequency-based and stacking classifier approach showed enhanced performance than other existing machine and deep learning techniques.
2. Individual machine learning classifiers are assessed using different parameters. A fusion of identical and a diverse set of machine learning classifiers with a frequency-based approach and stacking classifier maintains good efficiency, thus achieving state-of-the-art on posed datasets.
3. The proposed model is evaluated on both single and multi-pose datasets with fine-tuned parameters, making the model achieve better performance against pose variations.
4. The proposed model tries to reduce the errors by analyzing the pattern in the base classifier before making the final predictions.

The remainder of this paper includes six sections. Section 2 presents preliminaries about the methods used for the ensemble model. FER related works are outlined in Section 3. The proposed methodology is discussed in Section 4. Experiment results and analysis are summarized in Section 5. Finally, the concluding remarks, along with future work, are pointed out in Section 6.

2 Preliminaries

2.1 EfficientNet

Earlier deep learning models have reached a hardware memory limit issue; hence, an efficient model is required to improve the accuracy. Furthermore, the CNN's are computationally expensive compared to machine learning models, as Neural Networks heavily depend on the data, the problem considered, and the complex network required to solve it. But, the computational difficulty of these networks was solved using Graphical Processing Unit (GPU). Finally, the Google Research Brain team's latest model, the EfficientNet [43] (a variant of the CNN), achieved state-of-the-art accuracy, faster computation power, compactness, and overcame all previous deep learning models.

The ConvNets are scaled up to obtain better accuracy and efficiency. Hence, it is scaled up by depth, width, and resolution. Single dimension scaling models tend to achieve higher accuracy with larger depth, width, and resolution, but it has a limitation, the accuracy gain drops and saturates after reaching 80%. The EfficientNet model overcomes the drawback by compound scaling [43], i.e., by scaling three dimensions like width, depth, and resolution with a fixed ratio. This model starts from high quality and with a compact baseline model and scales up each of its dimensions uniformly with a fixed set of escalade coefficients. If the image's resolution is bigger in the compound scaling method, the network needs a more receptive field and more channels to capture other fine-grained patterns. In the proposed work, EfficientNet B0, a baseline model is utilized, and architecture details are given in Table 1.

Mobile Inverted Bottleneck Conv (MBConv) [14, 15, 26, 38], an inverted bottleneck Conv, is the main building block or main component of EfficientNet. It is also an inverted residual structure with an injection of Squeeze and Excite (SE) block, which has skip connections between thin bottleneck layers. The inverted residual blocks are efficient compared to classical residual networks, as propagating the gradient across multiplier layers is improved.

2.2 Stacking classifier

Stacking is a process of constructing classifier ensembles [1]. It is an ensemble learning technique that combines multiple classification models (machine learning classifiers) via meta classifier [27, 37]. It is an approach where several individual classifiers' outputs (decisions) are combined to classify new instances. The stacking process combines multiple classifiers [22, 29] to create high-level classifiers and produce improved performance. In the first level, the features are fed into the various base classifiers which, outputs a new decision. Later in a second level, a meta classifier decides the final prediction by considering the base classifiers' opinions and their prediction (output pattern) value [2]. Suppose if the base classifiers make some classification errors. In that case, the meta classifier can successfully learn the pattern and decide which prediction value to be considered for the final prediction. By doing so, the overall performance of the recognition system is improved. The bias and variance can be reduced using the stacking approach [1], as the combination of different ensemble components tries to learn from its errors. The stacking approach is flexible and powerful as compared to that of other ensemble methods.

Table 1 Architecture of EfficientNet B0 [43]

Stage i	Operator F_i	Resolution H_i x W_i	# Channels C_i	# Layers L_i
1	Conv3×3	224 × 224	32	1
2	MBConv1, k3×3	112 × 112	16	1
3	MBConv6, k3×3	112 × 112	24	2
4	MBConv6, k5×5	56 × 56	40	2
5	MBConv6, k3×3	28 × 28	80	3
6	MBConv6, k5×5	14 × 14	112	3
7	MBConv6, k5×5	14 × 14	192	4
8	MBConv6, k3×3	7 × 7	320	1
9	Conv1×1 & Pooling & FC	7 × 7	1280	1

3 Related works

Ensemble classifier [27], combines the decisions from the multiple classifiers instead of relying on a single classifier decision. Thus, it helps in improving the overall accuracy of the system through enhanced decision-making. The ensemble approach proved efficient in various studies [2, 32, 37] and gave the best accurate prediction. [49] proposed an ensemble CNN architecture and fused the probabilities of these CNN architectures using the probability-based fusion method. [53] utilized three state-of-the-art face detector modules, ensemble all three face detectors to improve the face detection, and followed the ensemble of various Deep Convolutional Neural Network (DCNN) with randomized initialization for classification of FER. An ensemble classifier based on the Dynamic Weight Majority Voting (DWMV) mechanism with an incremental learning property is proposed by [60] to learn various incoming expression patterns from images belonging to new expression classes. The combination of SURF with DWMV showed superior performance for FER.

According to [30], approaches like data augmentation and ensemble voting improve generalization performance. Hence, they proposed an ensemble of CNN architectures (VGG, Inception, ResNet) without utilizing additional training data or facial registration. This approach became a state-of-art method compared to previous CNN-based FER architectures. In contrast, an ensemble model is proposed by [23] with three distinct structured CNN subnets trained separately. The combination of all three ensemble subnets provided better performance results on FER 2013 dataset and obtained 5th rank in the competitions. Also, [9] proposed a Multi-Region Ensemble CNN (MRE-CNN) framework to detect the contribution of three different sub-regions of the human face. This framework rendered a remarkable performance by assigning the weights to these three networks and combining their final predictions. Finally, an ensemble of Deep CNN's with four different ensemble strategies like a seed, preprocessing, pretraining, and bagging is proposed by [34] to recognize facial expressions efficiently. The authors have also performed an extensive investigation on various aspects of ensemble generation and focused on the factors which influence classification accuracy. In contrast to the previous approach, this work focuses on saving the best model weights and extracting their intermediate features, feed them to the base classifiers of the ensemble model to recognize the facial expressions and improve the results efficiently.

4 Proposed methodology

The proposed work used for the recognition of facial expressions is briefly discussed in this section. The model consists of three stages: Pre-processing, Feature Extraction, and Classification. Finally, the facial images are mapped into respective expression classes using an ensemble approach as shown in Figs. 1 and 2. The proposed model's performance is evaluated on Oulu-CASIA and RaFD (multi-pose and only frontal pose images) datasets.

1. **Pre-processing and Data Augmentation**
   The static images are used to carry out the experiments. Every image present in the dataset is pre-processed. For the face detection, Paul Viola and Michael Jones [47, 48] Adaboost learning algorithm is used. This technique uses Haar-Like features and AdaBoost to train cascaded classifiers and detect the faces with a frontal view in lesser time [60]. Facial image contains a lot of unnecessary background information which

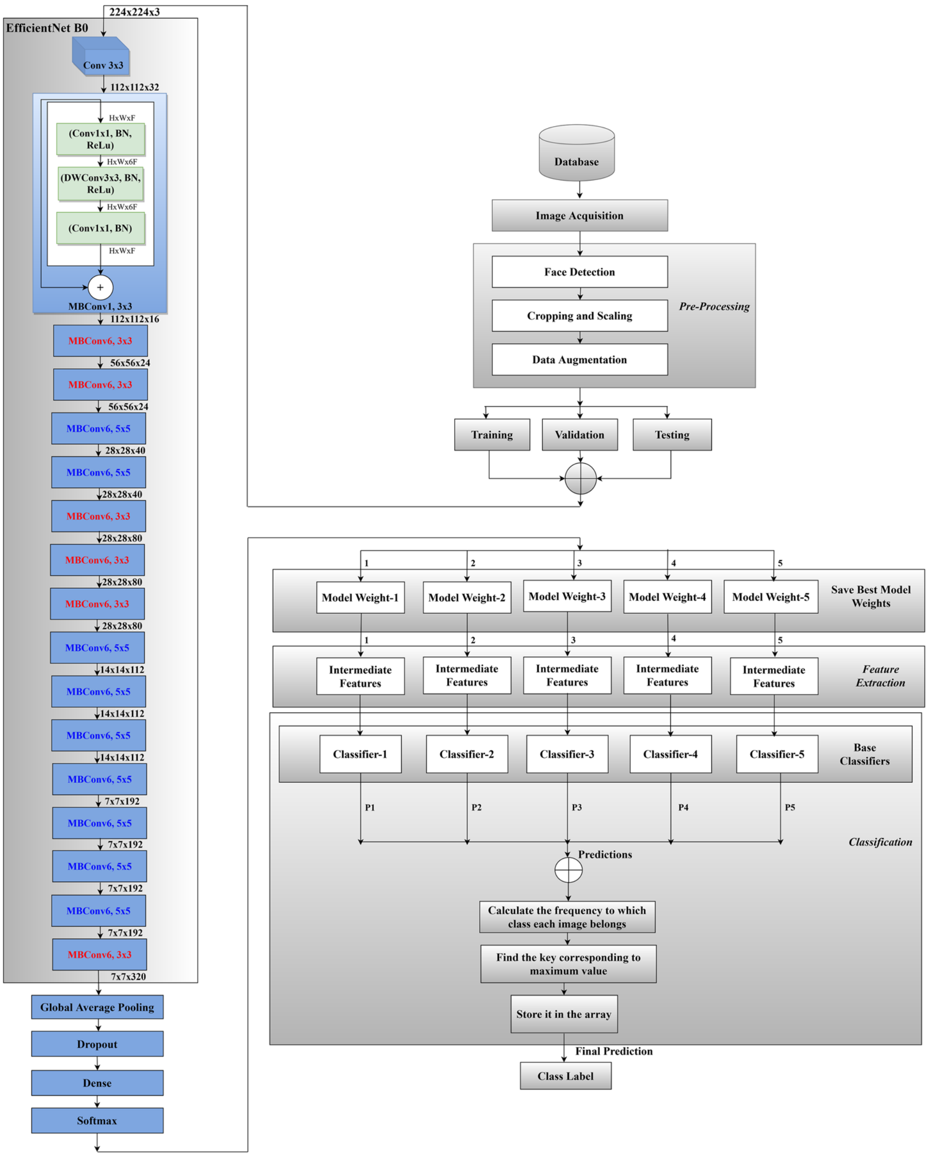


Fig. 1 Ensemble model architecture based on the frequency of votes (FV-EffNet)

is not useful for the classification of the expressions [25], hence this irrelevant information is cropped, and expression specific information is retained. This approach was successful on Oulu-CASIA and RaFD datasets with only a frontal pose. The Viola-Jones algorithm failed to detect the faces with pose variations [55]. Hence, a Multi-task Cascaded Convolutional Networks (MTCNN) is used to detect faces with multi-pose variations. MTCNN [55] is a deep cascaded architecture which exploits the innate correlation between the detection and alignment. This framework consists of three-stage multitask deep convolutional networks like P-Net, R-Net, and O-Net, designed

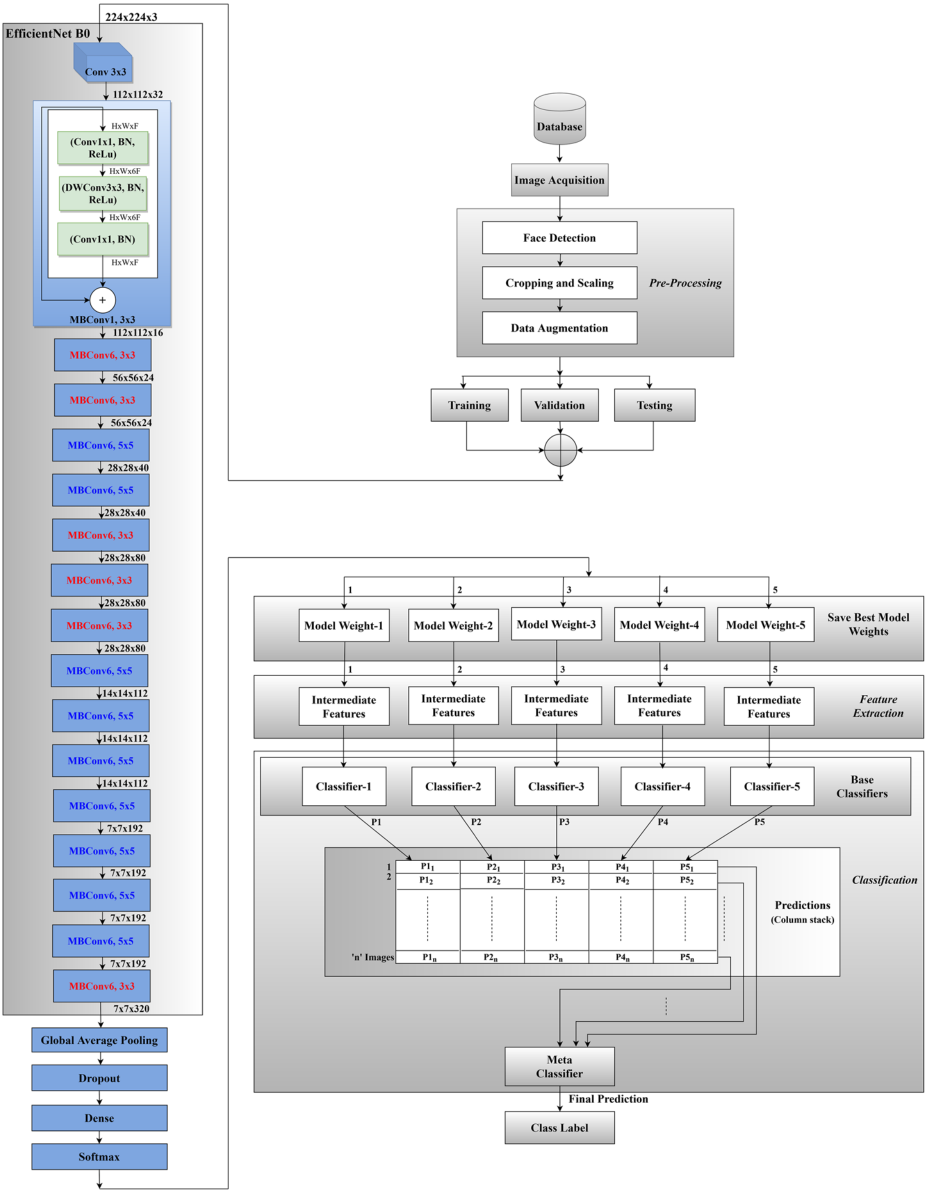


Fig. 2 Stacking classifier architecture (SC-EffNet)

to predict facial and landmark location in a coarse-to-fine manner. The MTCNN face detection technique was successful in detecting faces with a multi-pose variation on the RaFD dataset. All the images present in the dataset were resized into 224x224 pixel resolution and fed to the network for recognition.

The offline data augmentation is performed to improve the training samples. Data augmentation is a technique to virtually create extra training data by applying transformations to the input image, given the training data. In this work, horizontal and vertical flipping is applied on Oulu-CASIA and RaFD (frontal data) datasets. Data Augmentation has proven to be efficient in improving the generalization ability of deep learning models in various applications like image classification, speech recognition, and other areas. The huge complex designed network tend to over-fit on training data. Hence, to avoid this, it requires to feed a massive amount of data [23].

2. **Feature Extraction using deep learning Model**
   Feature extraction is a vital step in FER, and it is an aid to derive effective facial representations from the original facial image. There are two ways of extracting facial features: handcrafted features and the other using CNN architectures to extract auto-extracted features [45]. The extracted features play an essential role in minimizing the distance of intra-class variations and maximizing the distance between inter-class. The best classifier will also fail to achieve good performance if the extracted features are inadequate. The handcrafted feature extraction techniques like Local Binary Patterns (LBPs), Local Gabor Binary Patterns (LGBPs), Histograms of Oriented Gradients (HOG), and Scale Invariant Feature Transforms (SIFTs) have achieved great success in various fields with a small amount of training data. These low-level features are difficult to extract; tuning the features according to incoming face images and gathering discriminative information from these data is also tricky. These disadvantages present significant challenges in accurately recognizing expressions in real-life applications, as these data impose large inter-personal differences in appearance and capturing conditions. Hence, deep learning approaches cope up with these challenges and automatically discover multiple data representations and extract abstract concepts from a higher representation level. Thus, this was a reason for the breakthrough in recognition tasks.

   The EfficientNet model has achieved a state-of-the-art in image classification [43] and achieved high performance, and low computational cost [26]. In the proposed work, the EfficientNet B0 architecture has been used in the basic feature extraction process. Initially, the EfficientNet model is executed up to certain epochs, and all the weights are saved into a folder. Then, the model's five best weights are chosen based on the performance of the validation accuracy. Later, the best weights are loaded, and their respective intermediate features are extracted and fed into an ensemble of machine learning classifiers to improve the FER's efficiency.

3. **Classification**
   Combining the EfficientNet model's features and various machine learning classifiers proved to be advantageous [27]. This work presented two novel ensemble models, an EfficientNet model with machine learning classifiers using a frequency-based voting strategy (FV-EffNet) and an EfficientNet model with the stacking classifier (SC-EffNet) for classification. Various machine learning classifiers are evaluated empirically. The classifiers that generated the best results were chosen for further evaluation. The proposed approaches takes machine learning classifiers like Extremely Randomized Trees Classifier (Extra Trees classifier) [51], Random Forest (RF) [1], Decision Trees (DT) [36], K-Nearest Neighbors (KNN) [7], Multilayer Perceptron (MLP) [33], and Support Vector Machine (SVM) [28] as base classifiers.

   Every model's intermediate features are loaded individually and fed into each base classifier separately and evaluated to check with which intermediate features and base classifier (varying their parameters) the efficient result is obtained. This step is necessary as the features fed to the base classifier play a vital role in recognizing the input pattern and predicting the outcomes.

(a) Classification using a frequency-based voting approach: In FV-EffNet for each image, the predictions from five separate base classifiers are analyzed row by row (predicting the class to which each image belongs). The frequency (vote) is calculated using all five predictions for each image. Finally, the maximum vote from all five predictions is used to generate a key. The final key value is the final prediction (emotion class to which each image belongs) obtained from the combination of base classifiers, and it is stored in the array. The strategy is depicted in the Fig. 1.

   - Base classifiers: The EfficientNet B0 intermediate features are fed to various machine learning classifiers. The base classifiers like KNN, MLP, RF, Extra Trees, SVM on Oulu-CASIA, and Extra Trees, RF, DT, MLP, KNN on RaFD datasets are chosen to predict various expression classes efficiently.
   - Ensemble classifier: The predictions from the aggregation of base classifiers would outperform compared to predictions from a single model [27]. Hence, this is a reason behind the choice of an ensemble model to predict expression class. A frequency-based voting strategy combines the predictions from various machine learning base classifiers and makes a final decision. Hence, the ensemble model's output using a frequency-based algorithm would be a final class label predicted by most classifiers [32].

(b) Classification using stacking classifier: It is an ensemble learning technique that combines multiple classification models through meta-classifier [27]. Instead of bagging and boosting approach, stacking tries to learn how to combine the base classifiers (level 0 classifiers) rather than taking votes. A novel approach, where a combination of deep learning model and stacking classifier (SC-EffNet) is proposed and depicted in Fig. 2. The best features from the EfficientNet B0 model are fed to a stacking classifier [1, 44] for classification.

   The intermediate features are first fed to base classifiers (level 0 classifiers), containing diverse machine learning classifiers. Each base classifier is trained using training data. The predictions (output) from each base classifier are appended and stacked as a vector. These predictions are considered as a new dataset fed as input to the meta classifier (level 1 classifier) [1]. Later, the meta classifier is trained with this new dataset, and evaluation is done by performing cross-validation on test data. The meta classifier helps analyze the data pattern in a better way and helps to get accurate predictions. Finally, this classifier outputs the final prediction. One of the advantages of using a stacking classifier is that it decreases the risk of getting varied outputs from different machine learning classifiers. It clubs the results of all individual machine learning classifiers, analyzes the pattern, performs accurate predictions, and achieves good performance.

   - Base Classifiers: The machine learning classifiers like KNN, MLP, RF, Extra Trees, and SVM are used as a base classifier on the Oulu-CASIA dataset and Extra Trees, RF, DT, MLP, KNN on the RaFD dataset.
   - Meta Classifier: The Extra Tree classifier outperformed other machine learning classifiers and was chosen as a meta classifier for evaluation on the Oulu-CASIA dataset. During the RaFD dataset evaluation, the DT classifier proved to be efficient compared to other machine learning classifiers for the final prediction.

5 Experiment results and analysis

5.1 Dataset description

This section presents different datasets used for the evaluation of the proposed technique.

1. Oulu-CASIA: This facial expression database is a posed dataset [57]. It includes 480 image sequences elicited from 80 subjects in six different background settings using three illumination conditions: normal, weak, and dark. The cameras like Near-Infrared (NIR) and Visual (VIS) were used to capture the same facial expressions elicited by subjects. Each image sequence in the database begins with a neutral expression and ends with peak emotion labels. Each image's pixel resolution is 320*240. The images captured from visual cameras are used for the evaluation of the proposed methodology. In this experiment, peak expressions from the last three frames are chosen as training, testing, and validation data. A total of 236 images have been chosen for our experiments from 240 image segments after applying the viola jones face detection algorithm from all expression classes. Images are resized into 224*224 pixel resolution. Later, horizontal and vertical scaling augmentations are applied to increase the number of images. This dataset includes basic emotions like anger, disgust, fear, happiness, sadness, and surprise.
2. Radboud Faces Database (RaFD): The RaFD database [21, 39] contains images from 67 subjects collected using five camera angles and has five pose degrees 0, 45, 90, 135, 180 with three gaze directions (frontal, left and right views). The dataset includes expressions like anger, disgust, fear, happiness, sadness, surprise, contempt, and neutral. A total of 8040 images are present in this database, and each image pixel resolution is 681*1024. Two types of experimentation are carried out in this work using the RaFD database. One experiment is carried out with images with only a frontal pose, where the Viola-Jones algorithm is applied for face detection. The augmentation like horizontal and vertical scaling is used to increase the images with a frontal pose. The other experiment is carried out with the entire RaFD dataset, which includes all five pose angles. A total of 7974 facial images are detected using the MTCNN face detection approach. Data augmentation is not applied to this data with multi-pose angle, and all images are resized into 224*224 pixel resolution.

5.2 Implementation details

For training the network, this work has used a pre-trained network with pre-trained weights instead of training from scratch, and this is known as transfer learning. Transfer learning has proved to be effective in various computer vision applications [26]. This approach is applied to EfficientNet B0 and pre-trained on the ImageNet dataset, much broader than the facial images presented to the proposed models. The network weights are fine-tuned by the optimizer in the new training phase, allowing the model to adapt to our problem. The imported models have a lot of knowledge about the objects.

In the training phase, Adam optimizer is used to update the weights and reduce the learning rate by a factor of 10 in the event of stagnation ('patience=7'). The learning rate started at $1e^{-4}$, and the batch size is 10, and the number of epochs is fixed at 50. During training, an early stopping callback is used to control the overlearning of EfficientNet architecture. The model weights are saved using the model checkpoint. Rectified Linear Unit (ReLU) [46] activation function is used which transforms the linear input into non-linear data. ReLU is computed using formula $f(\varphi) = max(0, \varphi)$. With ReLU, the network becomes more efficient due to its sparse feature representation; it also helps in faster training, reduces computational complexity, and overcomes vanishing gradient problem. The softmax layer is used in the output layer of the EfficientNet model. It is used in multi-class classification problems [34] to estimate the testing sample's probabilities belonging to each class.

5.3 Experiment 1: Evaluation of efficientNet B0 model

In this experiment, the EfficientNet B0 model is used as a classifier for evaluating posed datasets, and the results are presented in Table 2. The model showed an accuracy of 97.28% and 98.53% with augmented data on Oulu-CASIA and RaFD (only 90deg frontal pose) datasets. Without data augmentation, the EfficientNet B0 model as a classifier achieved a performance of 93.72%, 95.10%, and 97.06% on Oulu-CASIA, RaFD datasets with frontal pose and Multi-pose variations, respectively.

5.4 Experiment 2: Proposed methodology

The entire dataset is split into training, testing, and validation set. Every epoch's weights are saved while monitoring the parameter val_acc using the model checkpoint. An early stopping callback is used to stop the EfficientNet model's training if there is no increase in the value of val_acc until patience=10 (10 iterations). Among all the saved weights, 'n' best weights are loaded where n=1 to 5. Their respective intermediate features are fed to the combination of machine learning base classifiers for classification as shown in Figs. 1 and 2.

The detailed experiment procedure outlining the entire flow of the proposed methodology is depicted in Fig. 3. First, according to the process, the model weights that achieved the best results are saved, and their respective intermediate features are loaded. Next, every machine learning classifier is adapted, and these classifiers' performance is verified on the saved 'n' models intermediate features. Finally, the classifier which gave the best results is chosen as a particular base classifier for that specific model. Similarly, the same procedure is followed for the rest of the 'n-1' models. The parameters of the machine learning classifiers are fine-tuned by varying the number of estimators, maximum depth of the tree, minimum samples of a leaf node, minimum sample split, maximum iterations of RF, DT, Extra Trees, SVM, MLP classifiers, and the nearest neighbor value of the KNN classifier. The parameters that showed the best performance results were eventually considered for individual classifiers. The Figs. 4 and 5 presents results obtained from the base classifiers when evaluated using identical set of machine learning classifiers.

5.4.1 Frequency-based voting strategy

Five best epochs intermediate features are fed to 'm' machine learning base classifiers where m=1 to 5. The predictions from five individual machine learning classifiers (base

Table 2 Experimental result obtained with efficientnet b0 architecture as a classifier

Model	Optimizer	Dataset	Test accuracy
EfficientNet B0	Adam	Oulu-CASIA (without augmentation)	93.72%
		Oulu-CASIA (with augmentation)	97.28%
		RaFD (90 deg) (without augmentation)	95.10%
		RaFD (90 deg) (with augmentation)	98.53%
		RaFD (Multi-Pose) (without augmentation)	97.06%

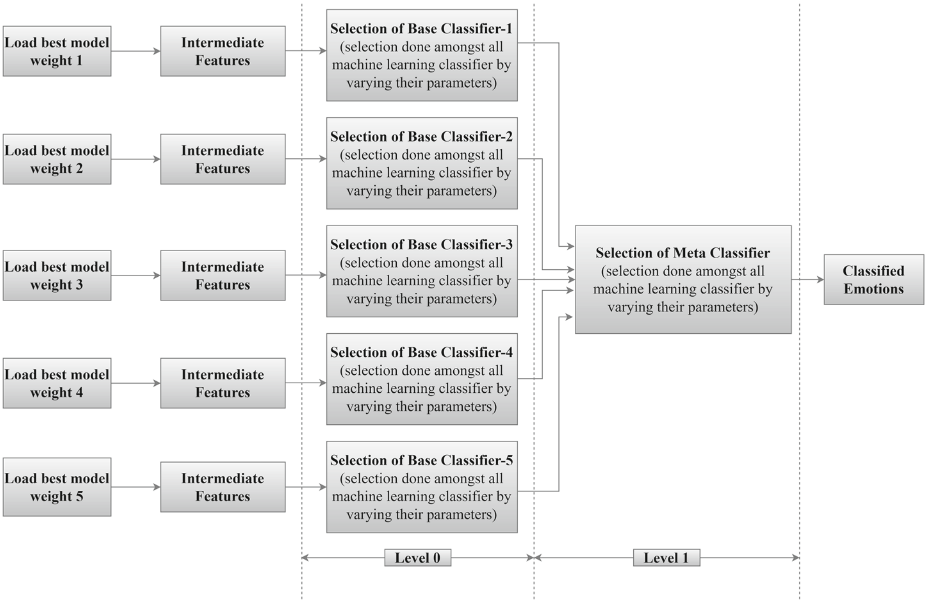

Load best model weight 1
Load best model weight 2
Load best model weight 3
Load best model weight 4
Load best model weight 5
Intermediate Features
Selection of Base Classifier-1 (selection done amongst all machine learning classifier by varying their parameters)
Selection of Base Classifier-2 (selection done amongst all machine learning classifier by varying their parameters)
Selection of Base Classifier-3 (selection done amongst all machine learning classifier by varying their parameters)
Selection of Base Classifier-4 (selection done amongst all machine learning classifier by varying their parameters)
Selection of Base Classifier-5 (selection done amongst all machine learning classifier by varying their parameters)
Selection of Meta Classifier (selection done amongst all machine learning classifier by varying their parameters)
Classified Emotions
Level 0
Level 1

Fig. 3 Detailed procedure outlining the flow of proposed methodology

classifiers) are appended and considered for further evaluation. The final class label is predicted by taking votes from most classifiers in the ensemble model. The experiment results obtained from the proposed approach are presented in Tables 3 and 4. The deep learning model (EfficientNet B0), combined with distinct classifiers using a frequency-based voting strategy, gave the best results of 98.71% and 98.56% on Oulu-CASIA and RaFD multi-pose datasets, respectively and also depicted in Figs. 7 and 8.

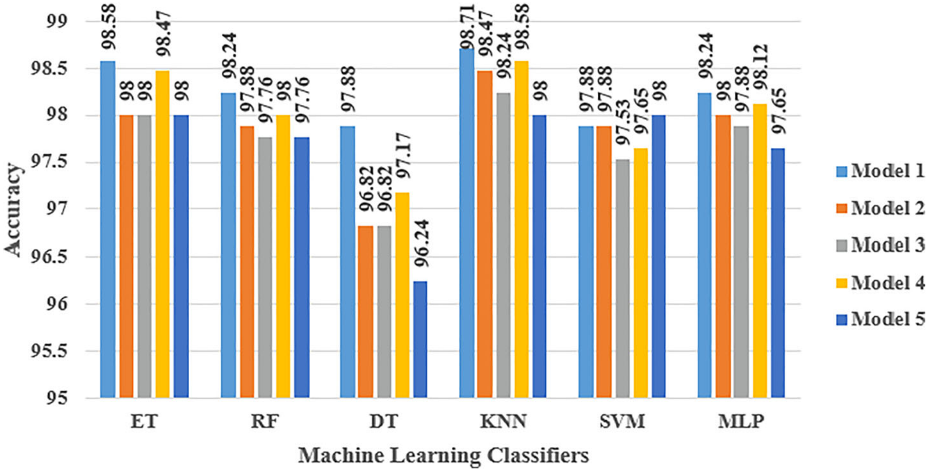


Fig. 4 Results obtained from base classifiers (level 0) on Oulu-CASIA dataset

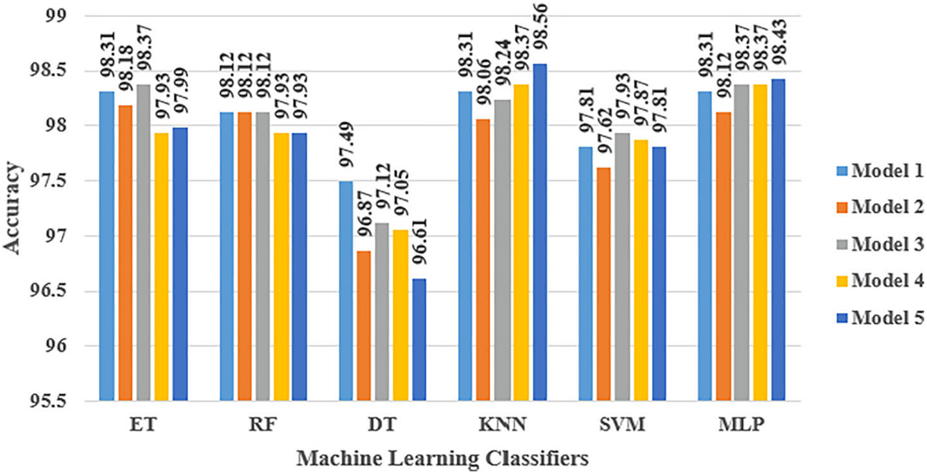


Fig. 5 Results obtained from base classifiers (level 0) on RaFD dataset (Multi-Pose)

5.4.2 Stacking classifier

Initially, five best epochs intermediate features are fed to ‘m’ machine learning classifiers (base classifiers) where m=1 to 5. After empirically testing each machine learning classifier, the classifiers that performed best are chosen as the base classifier. All the base classifiers are trained using training data, and their predictions are horizontally stacked and converted into vectors and fed to meta classifier to get the final prediction. Due to a lack of test data, the test set is subjected to cross-validation (cv=5). As a result, it establishes the robustness of the stacking strategy and the model’s generalizability. Each machine learning classifier (Extra Trees (ET), KNN, RF, DT, MLP, and SVM) is individually chosen for evaluation as a meta classifier. Further, as shown in Fig. 6, based on various meta classifiers’ performances, Extra Trees (ET) and DT classifiers proved to be efficient on Oulu-CASIA and RaFD datasets.

The results obtained when evaluating the identical base classifier and the same machine learning classifier chosen as meta-classifiers are depicted in Figs. 7 and 8 and also in Table 3. Similarly, the EfficientNet model with a distinct combination of machine learning classifiers and stacking classifiers are presented in Figs. 7 and 8 and Table 4. The accuracy of 98.35% and 98.06% is obtained with a stacking classifier approach on Oulu-CASIA and RaFD (multi-pose) datasets, respectively.

5.5 Observations

The confusion matrices obtained when evaluating a combination of individual machine learning classifiers are presented in Figs. 9 and 10. Also, the Figs. 11 and 12 depicts the confusion matrices obtained when evaluating a different combination of machine learning classifiers on Oulu-CASIA and RaFD (multi-pose) datasets, respectively. Thus, the observation is that every individual classifier contributes equivalently to classify the expressions into respective classes. For example, while observing the confusion matrix given in Fig. 9,

Table 3 The output of machine learning algorithms after fine-tuning individual classifier on Oulu-CASIA and RaFD datasets

Model	Base classifiers	Meta classifier	Oulu-CASIA			RaFD (Multi-Pose)		
			Output	Voting	Stacking classifier	Output	Voting	Stacking classifier
EfficientNet B0	ET 1	ET	98.58%	98.35%	98.47%	98.31%	98.24%	98.18%
	ET 2		98%			98.18%		
	ET 3		98%			98.37%		
	ET 4		98.47%			97.93%		
	ET 5		98%			97.99%		
	RF 1	RF	98.24%	98.12%	98%	98.12%	98.24%	98.18%
	RF 2		97.88%			98.12%		
	RF 3		97.76%			98.12%		
	RF 4		98%			97.93%		
	RF 5		97.76%			97.93%		
	DT 1	DT	97.88%	98.24%	97.65%	97.49%	98.31%	96.74%
	DT 2		96.82%			96.87%		
	DT 3		96.82%			97.12%		
	DT 4		97.17%			97.05%		
	DT 5		96.24%			96.61%		
	KNN 1	KNN	98.71%	98.47%	98.24%	98.31%	98.18%	97.93%
	KNN 2		98.47%			98.06%		
	KNN 3		98.24%			98.24%		
	KNN 4		98.58%			98.37%		
	KNN 5		98%			98.56%		
	SVM 1	SVM	97.88%	98.35%	97.83%	97.81%	98.24%	96.31%
	SVM 2		97.88%			97.62%		
	SVM 3		97.53%			97.93%		
	SVM 4		97.65%			97.87%		
	SVM 5		98%			97.81%		
	MLP 1	MLP	98.24%	98.47%	98.24%	98.31%	98.37%	97.93%
	MLP 2		98%			98.12%		
	MLP 3		97.88%			98.37%		
	MLP 4		98.12%			98.37%		
	MLP 5		97.65%			98.43%		

*Meta classifier is considered only for the evaluation of stacking classifier approach

the Extra Tree, RF, and DT classifiers predict 108 images correctly into anger expression classes and does one misclassification into a sadness expression class. Similarly, the KNN classifier predicts 107 images correctly into anger expression class and does two misclassifications. Whereas SVM and MLP classifier precisely classifies all 109 images into proper

Table 4 The output of distinct machine learning algorithms after fine-tuning on Oulu-CASIA and RaFD datasets

Datasets	Base classifiers	Meta classifier	Output	Voting	Stacking classifier
Oulu-CASIA	KNN		98.71%	98.71%	98.35%
	MLP		98%		
	RF	ET	97.76%		
	ET		98.47%		
	SVM		97.88%		
RaFD (Multi Pose)	ET		98.31%	98.56%	98.06%
	RF		98.12%		
	DT	DT	97.12%		
	MLP		98.37%		
	KNN		98.56%		

*Meta classifier is considered only for the evaluation of stacking classifier approach

expression classes. Thus, the conceptual lesson to be learned from this proposed work is that every individual classifier is responsible for categorizing expressions into an appropriate emotion class.

The best features fed into an ensemble of machine learning classifiers showed enhanced performance on the frontal pose and multi-pose datasets. By fusing the outputs of base classifiers and providing them to the higher-level classifier, we try to reduce the errors by ana-

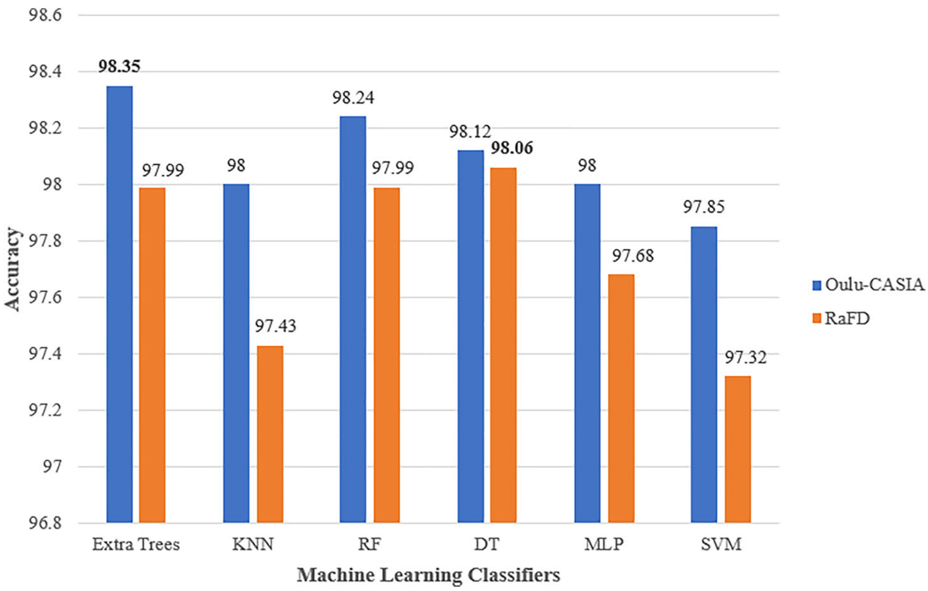


Fig. 6 Selection of meta classifier (level 1) for evaluation of distinct set of base classifiers in stacking classifier approach

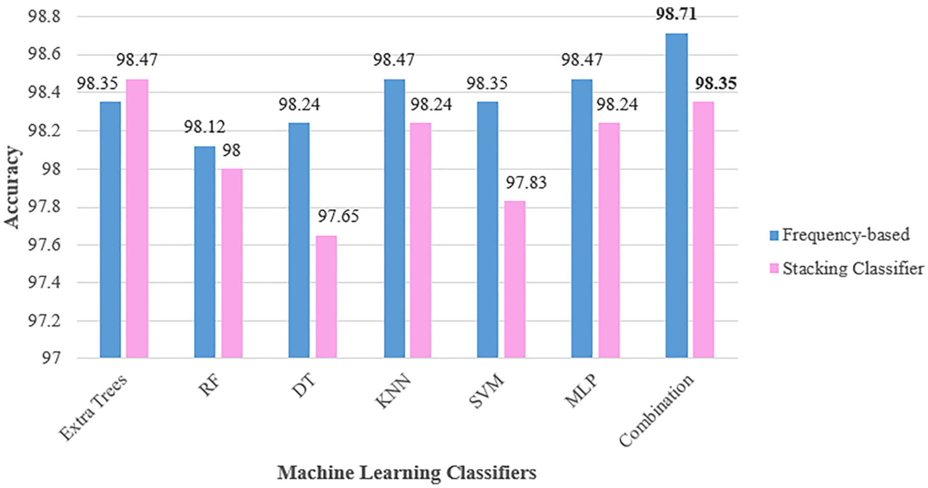


Fig. 7 Results obtained when evaluating the frequency-based and stacking classifier approaches on Oulu-CASIA dataset

lyzing the pattern before making the final predictions. Thus, it suggests that a combination of classifiers using the stacking approach is better than selecting the best single classifier for classification. It will help improve the system's efficiency and overcome the mistakes made in the previous classification level. The ensemble of deep learning and machine

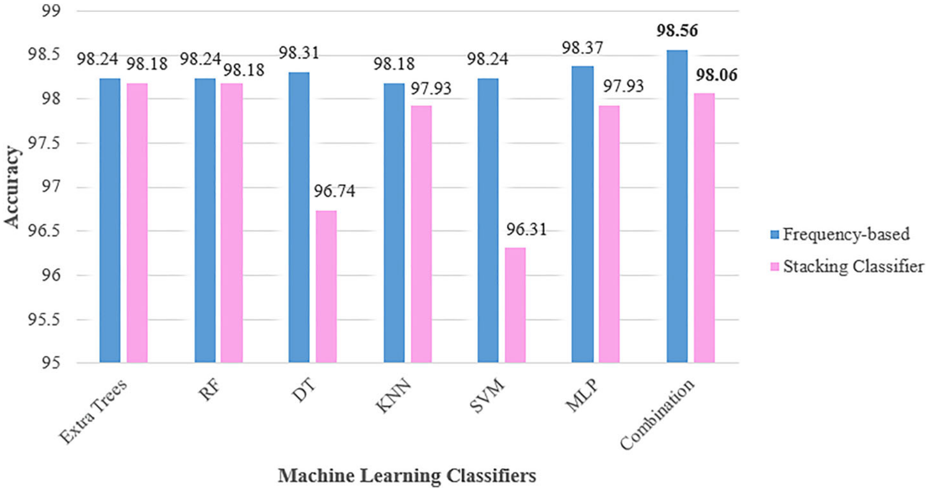


Fig. 8 Results obtained when evaluating the frequency-based and stacking classifier approaches on RaFD dataset (Multi-Pose)

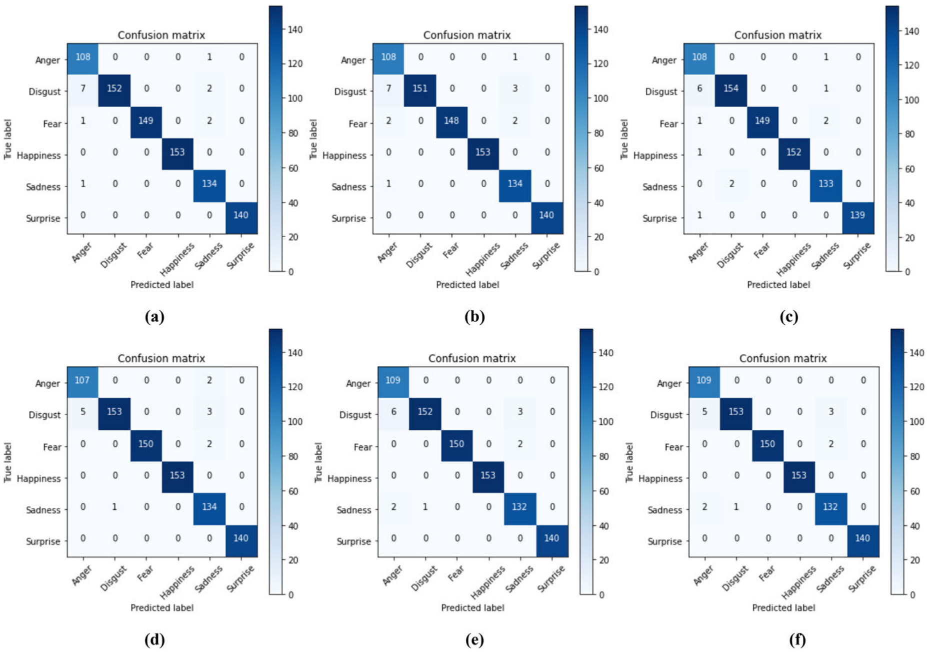


Fig. 9 Confusion Matrices obtained from machine learning classifiers on Oulu-CASIA dataset ((a) Extra Tree Classifier (b) Random Forest (c) Decision Trees (d) k-Nearest Neighbors (e) Support Vector Machine (f) Multi-Layer Perceptron)

learning techniques performs better than the earlier methods, thus showing the state-of-the-art on Oulu-CASIA and RaFD datasets. The proposed approach is robust against pose variations and involves multiple processing stages. However, majority voting predominately aids in enhancing the effectiveness of the system.

5.6 Experiment analysis and comparisons

The experiment results obtained with previous FER studies are presented in Table 5 and compared with the proposed approach. With the proposed methodology, the accuracies of 98.71% and 98.56% using frequency-based strategy and 98.35% and 98.06% using a stacking classifier approach are obtained on Oulu-CASIA and RaFD multi-pose datasets, respectively. The performance of the proposed approach is compared with other machine learning methods, CNN-based methods, and state-of-the-art results on the Oulu-CASIA and RaFD datasets. As observed in Table 5, the proposed model achieves better results than other methods on these benchmark facial expression databases.

Using RaFD datasets with a frontal pose, the proposed model outperforms [40] and [13] by 0.83% and 1.42%, respectively. The authors used a pyramid histogram of an orientated gradient for feature extraction with KNN classification in [39] and attained 100% accuracy. With an ensemble model, the proposed method likewise obtained 100% accuracy.

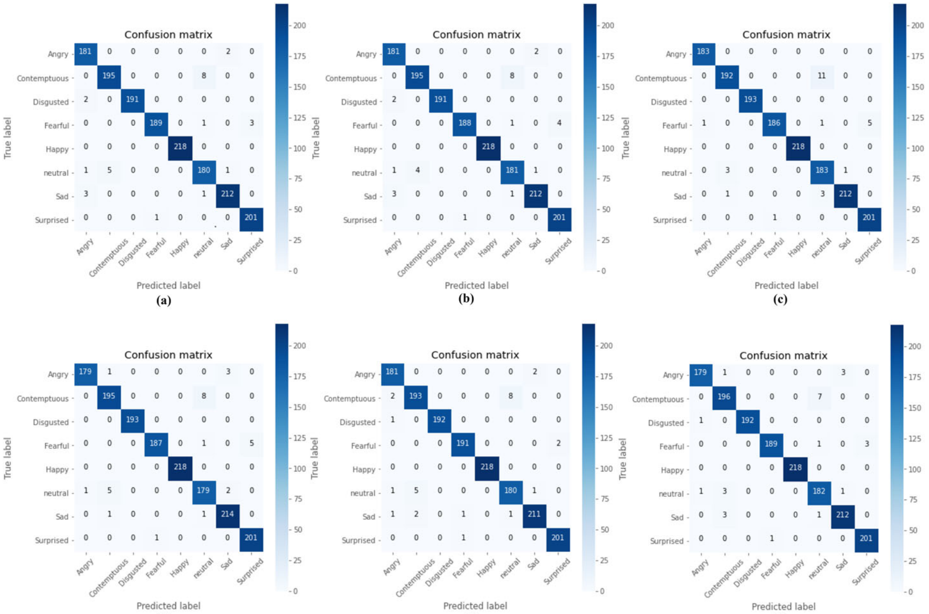


Fig. 10 Confusion Matrices obtained from machine learning classifiers on RaFD (Multi-Pose) dataset ((a) Extra Tree Classifier (b) Random Forest (c) Decision Trees (d) k-Nearest Neighbors (e) Support Vector Machine (f) Multi-Layer Perceptron)

Additionally, eight expression classes were considered in the proposed work, which improved the accuracy rate to 2.29%, outperforming [50]. The authors in [50] avoided the contemptuous class and employed seven expression classes. On the Oulu-CASIA dataset,

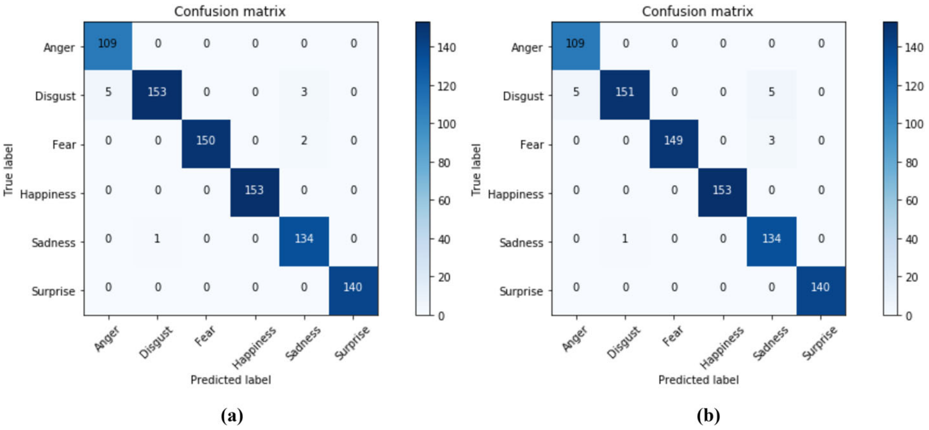
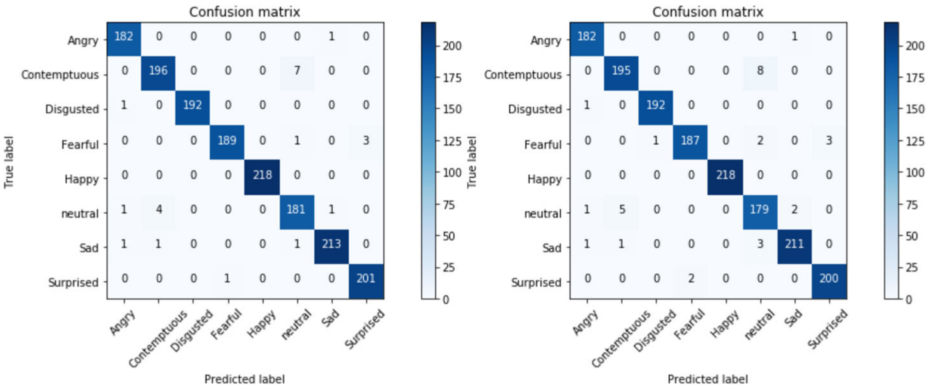


Fig. 11 Confusion Matrix obtained from the combination of various machine learning classifiers on Oulu-CASIA dataset ((a) Majority Voting (b) Stacking Classifier)



(a) **(b)**

Fig. 12 Confusion Matrix obtained from the combination of various machine learning classifiers on RaFD (Multi-Pose) dataset ((a) Majority Voting (b) Stacking Classifier)

the proposed technique had an enhanced accuracy of 10.71% compared to [52], which extracted the expressive component through a deexpression mechanism.

Table 6 presents the experiment results evaluated using various other performance metrics. When using the RaFD dataset with a frontal pose, the proposed model performs better in terms of precision, recall, and F1-score than [3]. Before making the final predictions, the proposed model examines the pattern in the base classifier to minimize the errors. Hence, the results showed better performance compared to previous studies making the proposed system robust against pose variations.

6 Conclusion

An ensemble model with a frequency-based voting approach (FV-EffNet) and a stacking classifier approach (SC-EffNet) is adopted to classify the expressions into respective classes. Combining multiple base classifiers induces the higher level classifiers to learn the pattern and thus help the ensemble model make accurate predictions rather than selecting a single classifier for classification. In the proposed methodology, even though both the ensemble models gave the best results, majority voting predominantly helped improve the system's performance.

The following conclusions are drawn from the experimental results: (1) The selection of best model weights and features extracted from EfficientNet showed better performance of 98.71% and 98.35% on Oulu-CASIA, and 98.56% and 98.06% on RaFD (multi-pose) datasets using frequency-based and stacking classifier approach compared to a baseline model. (3) An ensemble model with a frequency-based approach showed an improvement of 10.71% on the Oulu-CASIA dataset compared to [6] and 2.29% on the RaFD multi-pose dataset achieving the best performance than [50]. (4) The stacking classifier approach showed an improved efficiency by 10.35% and 1.79% on Oulu-CASIA and RaFD datasets,

Table 5 Comparison with previous approaches on Oulu-CASIA and RaFD datasets

Dataset	Method	Experimental settings	Accuracy
Oulu CASIA	DTAGN (Joint) [18]	Sequence-based	81.46%
	STM-ExpLet [24]	Sequence-based	74.59%
	DFSN-I [45]	Sequence-based	87.50%
	PHRNN-MSCNN [56]	Sequence-based	86.25%
	Microexpnet [4]	Image-based	85.83%
	PPDN [58]	Image-based	84.59%
	DeRL [52]	Image-based	88%
	FN2EN [6]	Image-based	87.71%
	EfficientNet B0 (Baseline)	Image-based	97.28%
	Proposed methodology using majority voting (FV-EffNet)	Image-based	**98.71%**
	Proposed methodology using stacking classifier (SC-EffNet)	Image-based	**98.35%**
RaFD (Frontal pose)	18-layered Conv-Deconv [42]	Image-based	93.41%
	CNN [10]	Image-based	93.33%
	MDSTFN [40]	Image-based	99.17%
	CNN [12]	Image-based	95%
	Weakly supervised learning [13]	Image-based	98.58%
	Hybrid-based AFER [51]	Image-based	96.16%
	Metric Learning[17]	Image-based	95.95%
	BAE-BNN-3 [41]	Image-based	96.93%
	HOG-SVM [3]	Image-based	98.2%
	EfficientNet B0 (Baseline)	Image-based	98.53%
	Proposed methodology using majority voting (FV-EffNet)	Image-based	**100%**
	Proposed methodology using stacking classifier (SC-EffNet)	Image-based	**100%**
RaFD (Multi Pose)	MP-AdaBoost [16]	Image-based	82.68%
	SURF boosting [31]	Image-based	90.64%
	W-CR-AFM [50]	Image-based	96.27%
	PHOG-KNN [39]	Image-based	100% (90deg), 96.7% (45deg to the right) and 98.1% (45deg to the left)
	Semi-supervised DBN [20]	Image-based	91.95% (135deg), 94.50% (90deg) and 92.75% (45deg)
	EfficientNet B0 (Baseline)	Image-based	97.06%
	Proposed methodology using majority voting (FV-EffNet)	Image-based	**98.56%**
	Proposed methodology using stacking classifier (SC-EffNet)	Image-based	**98.06%**

Table 6 Experiment results using other performance metrics

Dataset	Method	Experimental settings	Precision	Recall	F1-score
RaFD (Frontal pose)	HOG-SVM [3]	Image-based	93%	92.9%	92.9%
RaFD (Frontal pose)	Proposed methodology	Image-based	100%	100%	100%
RaFD (Multi pose)	Proposed methodology (FV-EffNet)	Image-based	98.53%	98.54%	98.53%
RaFD (Multi pose)	Proposed methodology (SC-EffNet)	Image-based	98.02%	98.04%	98.02%
Oulu-CASIA	Proposed methodology (FV-EffNet)	Image-based	98.56%	98.83%	98.67%
Oulu-CASIA	Proposed methodology (SC-EffNet)	Image-based	98.22%	98.51%	98.33%

respectively. This method decreases the risk of getting varying results from various machine learning classifiers and reduces bias and variance. The cross-validation performed on the test set proved the model's robustness and generalizability. (5) The proposed method with multi-stage processing showed better results with pose variations. The future work would be to evaluate the proposed approach on the spontaneous and in-the-wild databases and build a fully automated system that could be feasible for deploying real-world applications.

Author Contributions Conceptualization, Literature Review, original draft preparation: Rashmi Adyapady R.,Supervision, guidance and review: Annappa B.

Funding No funds, grants, or other support was received.

Availability of data and material The datasets discussed in the manuscript are publicly available for research purposes.

Code Availability We would like to share the code once the publication is authorised.

Declarations

Ethics approval The authors adhere to all the ethics declaration and guarantee that no discrepancies have occurred in the manuscript.

Consent to participate Not applicable.

Consent for Publication The author and co-author of the manuscript provide consent for publication in Multimedia Tools and Applications.

Conflict of Interests The authors declare that they have no conflicts of interest.

References

1. Aggarwal CC (2015) Data classification. In: Data mining, Springer, pp 285–344
2. Álvarez A, Sierra B, Arruti A, López-Gil JM, Garay-Vitoria N (2016) Classifier subset selection for the stacked generalization method applied to emotion recognition in speech. Sensors 16(1):21

3. Carcagnì P, Del Coco M, Leo M, Distante C (2015) Facial expression recognition and histograms of oriented gradients: a comprehensive study. SpringerPlus 4(1):645
4. Cugu I, Sener E, Akbas E (2019) Microexpnet: an extremely small and fast model for expression recognition from face images. In: 2019 Ninth international conference on image processing theory, tools and applications (IPTA), IEEE, pp 1–6
5. Dailey MN, Joyce C, Lyons MJ, Kamachi M, Ishi H, Gyoba J, Cottrell GW (2010) Evidence and a computational explanation of cultural differences in facial expression recognition. Emotion 10(6):874–893
6. Ding H, Zhou SK, Chellappa R (2017) Facenet2expnet: Regularizing a deep face recognition net for expression recognition. In: 2017 12th IEEE international conference on automatic face & gesture recognition (FG 2017), IEEE, pp 118–126
7. Dino HI, Abdulrazzaq MB (2019) Facial expression classification based on svm, knn and mlp classifiers. In: 2019 International conference on advanced science and engineering (ICOASE), IEEE, pp 70–75
8. Ekman P (2009) Darwin's contributions to our understanding of emotional expressions. Philosophical Transactions of the Royal Society B: Biological Sciences 364(1535):3449–3451
9. Fan Y, Lam JC, Li VO (2018) Multi-region ensemble convolutional neural network for facial expression recognition. In: International conference on artificial neural networks, Springer, pp 84–94
10. Fathallah A, Abdi L, Douik A (2017) Facial expression recognition via deep learning. In: 2017 IEEE/ACS 14Th international conference on computer systems and applications (AICCSA), IEEE, pp 745–750
11. Ghimire D, Lee J (2013) Geometric feature-based facial expression recognition in image sequences using multi-class adaboost and support vector machines. Sensors 13(6):7714–7734
12. González-Hernández F, Zatarain-Cabada R, Barrón-Estrada ML, Rodríguez-Rangel H (2018) Recognition of learning-centered emotions using a convolutional neural network. J Intell Fuzzy Syst 34(5):3325–3336
13. Happy S, Dantcheva A, Bremond F (2019) A weakly supervised learning technique for classifying facial expressions. Pattern Recogn Lett 128:162–168
14. Howard A, Sandler M, Chu G, Chen LC, Chen B, Tan M, Wang W, Zhu Y, Pang R, Vasudevan V et al (2019) Searching for mobilenetv3. In: Proceedings of the IEEE international conference on computer vision, pp 1314–1324
15. Howard AG, Zhu M, Chen B, Kalenichenko D, Wang W, Weyand T, Andreetto M, Adam H (2017) Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv:170404861
16. Jiang B, Jia K (2013) Semi-supervised facial expression recognition algorithm on the condition of multipose. J Inf Hid Multimed Signal Process 4:138–146
17. Jiang B, Jia K (2016) Robust facial expression recognition algorithm based on local metric learning. Journal of Electronic Imaging 25(1):013022
18. Jung H, Lee S, Yim J, Park S, Kim J (2015) Joint fine-tuning in deep neural networks for facial expression recognition. In: Proceedings of the IEEE international conference on computer vision, pp 2983–2991
19. Ko B (2018) A brief review of facial emotion recognition based on visual information. Sensors 18(2):401
20. Kurup AR, Ajith M, Ramón MM (2019) Semi-supervised facial expression recognition using reduced spatial features and deep belief networks. Neurocomputing 367:188–197
21. Langner O, Dotsch R, Bijlstra G, Wigboldus DH, Hawk ST, Van Knippenberg A (2010) Presentation and validation of the radboud faces database. Cognition and Emotion 24(8):1377–1388
22. Li W, Zou L (2017) Classifier stacking for native language identification. In: Proceedings of the 12th workshop on innovative use of NLP for building educational applications, pp 390–397
23. Liu K, Zhang M, Pan Z (2016) Facial expression recognition with cnn ensemble. In: 2016 International conference on cyberworlds (CW), IEEE, pp 163–166
24. Liu M, Shan S, Wang R, Chen X (2016) Learning expressionlets via universal manifold model for dynamic facial expression recognition. IEEE Trans Image Process 25(12):5920–5932
25. Lopes AT, de Aguiar E, De Souza AF, Oliveira-Santos T (2017) Facial expression recognition with convolutional neural networks: coping with few data and the training sample order. Pattern Recogn 61:610–628
26. Luz EJdS, Silva PL, Silva R, Silva L, Moreira G, Menotti D (2020) Towards an effective and efficient deep learning model for covid-19 patterns detection in x-ray images. CoRR
27. Malmasi S, Dras M (2018) Native language identification with classifier stacking and ensembles. Computational Linguistics 44(3):403–446
28. Michel P, El Kaliouby R (2003) Real time facial expression recognition in video using support vector machines. In: Proceedings of the 5th international conference on Multimodal interfaces, pp 258–264
29. Mihalcea R (2002) Classifier stacking and voting for text filtering. In the proceedings of Eleventh Text Retrieval Conference (TREC), pp. 696–701

30. Pramerdorfer C, Kampel M (2016) Facial expression recognition using convolutional neural networks: state of the art. arXiv:161202903
31. Rao Q, Qu X, Mao Q, Zhan Y (2015) Multi-pose facial expression recognition based on surf boosting. In: 2015 international conference on affective computing and intelligent interaction (ACII), IEEE, pp 630–635
32. Rao RS, Vaishnavi T, Pais AR (2019) Phishdump: a multi-model ensemble based technique for the detection of phishing sites in mobile devices. Pervasive and Mobile Computing 60:101084
33. Rashid TA (2016) Convolutional neural networks based method for improving facial expression recognition. In: The international symposium on intelligent systems technologies and applications, Springer, pp 73–84
34. Renda A, Barsacchi M, Bechini A, Marcelloni F (2019) Comparing ensemble strategies for deep learning: an application to facial expression recognition. Expert Syst Appl 136:1–11
35. Russell JA (1991) Culture and the categorization of emotions. Psychological Bulletin 110(3):426–450
36. Safavian SR, Landgrebe D (1991) A survey of decision tree classifier methodology. IEEE Trans Syst Man Cybern 21(3):660–674
37. Sakkis G, Androutsopoulos I, Paliouras G, Karkaletsis V, Spyropoulos CD, Stamatopoulos P (2001) Stacking classifiers for anti-spam filtering of e-mail. arXiv:cs/0106040
38. Sandler M, Howard A, Zhu M, Zhmoginov A, Chen LC (2018) Mobilenetv2: Inverted residuals and linear bottlenecks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4510–4520
39. Shokrani S, Moallem P, Habibi M (2014) Facial emotion recognition method based on pyramid histogram of oriented gradient over three direction of head. In: 2014 4Th international conference on computer and knowledge engineering (ICCKE), IEEE, pp 215–220
40. Sun N, Li Q, Huan R, Liu J, Han G (2019) Deep spatial-temporal feature fusion for facial expression recognition in static images. Pattern Recogn Lett 119:49–61
41. Sun W, Zhao H, Jin Z (2017) An efficient unconstrained facial expression recognition algorithm based on stack binarized auto-encoders and binarized neural networks. Neurocomputing 267:385–395
42. Sun W, Zhao H, Jin Z (2018) A complementary facial representation extracting method based on deep learning. Neurocomputing 306:246–259
43. Tan M, Le QV (2019) Efficientnet: Rethinking model scaling for convolutional neural networks. arXiv:190511946
44. Tang J, Alelyani S, Liu H (2015) Data classification: Algorithms and applications. In: Data mining and knowledge discovery series, CRC Press, pp 498–500
45. Tang Y, Zhang XM, Wang H (2018) Geometric-convolutional feature fusion based on learning propagation for facial expression recognition. IEEE Access 6:42532–42540
46. Verma M, Vipparthi SK, Singh G, Murala S (2019) Learnet: Dynamic imaging network for micro expression recognition. IEEE Trans Image Process 29:1618–1627
47. Viola P, Jones M (2001) Rapid object detection using a boosted cascade of simple features. In: Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001, vol 1. IEEE, pp 511–518
48. Viola P, Jones MJ (2004) Robust real-time face detection. Int J Comput Vis 57(2):137–154
49. Wen G, Hou Z, Li H, Li D, Jiang L, Xun E (2017) Ensemble of deep neural networks with probability-based fusion for facial expression recognition. Cognitive Comput 9(5):597–610
50. Wu BF, Lin CH (2018) Adaptive feature mapping for customizing deep learning based facial expression recognition model. IEEE Access 6:12451–12461
51. Yaddaden Y, Adda M, Bouzouane A, Gaboury S, Bouchard B (2018) Hybrid-based facial expression recognition approach for human-computer interaction. In: 2018 IEEE 20Th international workshop on multimedia signal processing (MMSP), IEEE, pp 1–6
52. Yang H, Ciftci U, Yin L (2018) Facial expression recognition by de-expression residue learning. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2168–2177
53. 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, pp 435–442
54. Zhang F, Zhang T, Mao Q, Xu C (2018) Joint pose and expression modeling for facial expression recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3359–3368
55. Zhang K, Zhang Z, Li Z, Qiao Y (2016) Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Process Lett 23(10):1499–1503
56. Zhang K, Huang Y, Du Y, Wang L (2017) Facial expression recognition based on deep evolutional spatial-temporal networks. IEEE Trans Image Process 26(9):4193–4203

57. Zhao G, Huang X, Taini M, Li SZ, PietikäInen M (2011) Facial expression recognition from near-infrared videos. Image Vis Comput 29(9):607–619
58. Zhao X, Liang X, Liu L, Li T, Han Y, Vasconcelos N, Yan S (2016) Peak-piloted deep network for facial expression recognition. In: European conference on computer vision, Springer, pp 425–442
59. Zhong L, Liu Q, Yang P, Huang J, Metaxas DN (2014) Learning multiscale active facial patches for expression analysis. IEEE Trans Cybern 45(8):1499–1510
60. Zia MS, Hussain M, Jaffar MA (2018) A novel spontaneous facial expression recognition using dynamically weighted majority voting based ensemble classifier. Multimed Tools Appl 77(19):25537–25567

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.