

Ensemble-based face expression recognition approach for image sentiment analysis

Ervin Gubin Moun¹, Chai Chuan Woo¹, Maisarah Mohd Sufian¹, Chin Kim On¹,
Jamal Ahmad Dargham²

¹Faculty of Computing and Informatics, Universiti Malaysia Sabah, Kota Kinabalu, Malaysia

²Faculty of Engineering, Universiti Malaysia Sabah, Kota Kinabalu, Malaysia

Article Info

Article history:

Received Jun 3, 2021

Revised Dec 14, 2021

Accepted Jan 2, 2022

Keywords:

Classification

Convolutional neural network

Ensemble

Facial expression recognition

Image sentiment analysis

InceptionV3

ResNet50

ABSTRACT

Sentiment analysis based on images is an evolving area of study. Developing a reliable facial expression recognition (FER) device remains a difficult challenge as recognizing emotional feelings reflected in an image is dependent on a diverse set of factors. This paper presented an ensemble-based model for FER that incorporates multiple classification models: i) customized convolutional neural network (CNN), ii) ResNet50, and iii) InceptionV3. The model averaging ensemble classifier method is used to ensemble the predictions from the three models. Subsequently, the proposed FER model is trained and tested on a dataset with an uncontrolled environment (FER-2013 dataset). The experiment demonstrated that ensembling multiple classifiers outperformed all single classifiers in classifying positive and neutral expressions (91.7%, 81.7% and 76.5% accuracy rate for happy, surprise, and neutral, respectively). However, when classifying disgust, anger, and sadness, the ResNet50 model alone is the better choice. Although the Custom CNN performs the best in classifying fear expression (55.7% accuracy), the proposed FER model can still classify fear expression with comparable performance (52.8% accuracy). This paper demonstrated the potential of using the ensemble-based method to enhance the performance of FER. As a result, the proposed FER model has shown a 72.3% accuracy rate.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Ervin Gubin Moun¹,

Faculty of Computing and Informatics, Universiti Malaysia Sabah

Kota Kinabalu 88400, Sabah, Malaysia

Email: ervin@ums.edu.my

1. INTRODUCTION

Sentiment analysis (SA) is a technique for determining consumer opinions about a specific topic, product, or issue in a more productive manner. It can be conducted with various modalities by taking text, image, audio, and video inputs [1]. SA based on textual data from various social media sources provides a simple way for businesses to gather customer feedback and further develop their products based on current market demands or trends. Machine learning has been used as a data processing technique to solve a wide range of problems in a variety of fields, including face recognition [2]–[4] and facial expression recognition. The facial expression is the most noticeable expression to distinguish a human emotion. According to Patel *et al.* [5], facial expressions can be classified into seven universal classes; i) happiness, ii) surprise, iii) contempt, iv) sadness, v) anger, vi) disgust, and vii) fear. In this paper, facial expression is used to recognize a person's emotional sentiment.

While many effective recognition systems have been implemented in the past, the recognition rate is not generally satisfied due to inherent disadvantages such as light, pose changes, noise, and occlusion [6]. Numerous facial expression recognition (FER) research reports have been published recently, but not much research is conducted on the fusion method of the FER model. Most of them used only one feature representation or fusion of feature representation to infer an expression [7], [8]. However, the performance of the ensemble-based FER model still has many rooms that can be investigated. Hence, the paper's first objective is to develop a FER model using an ensemble of existing FER models. Furthermore, most of the reviewed FER models were trained and tested with images under a controlled environment (image without occlusion, noise, and captured at a perfect angle), which contrasts with the real-world application [9], [10]. Given the unreliability of existing FER models in real-world applications, the second objective of this paper is to evaluate the proposed FER model's results using the FER-2013 dataset, a facial expression database that is highly heterogeneous and diverse and displaying emotion in its natural state [11].

This research adopted the same approach by Liu *et al.* [12], where they ensembled their own structured convolutional neural network (CNN) for FER. However, instead of structuring each of the ensemble members from scratch, the present study makes use of the pre-existing CNN architectures known for their promising results in the previous studies; residual network (ResNet) [13] and inception [14]. Furthermore, CNN also shows excellent success in extracting facial features [15], [16]. Currently, local binary pattern (LBP) is the state-of-art in the feature representation for FER [17]. However, it is well-known that this method is sensitive to abrupt changes in illumination. This paper compares three feature representations, namely: i) LBP-based feature representation, ii) CNN-based feature representation, and iii) LBP+CNN-based feature representation, in an attempt to find the best feature representation for the proposed ensemble-based model. Finally, this paper has the following contributions:

- a. Performed preliminary experiments to find out the best feature representation for the proposed ensemble-based FER model
- b. Proposed an ensemble-based FER model to address FER in SA and empirically evaluated its performance
- c. Emphasized the unreliability of the existing FER models toward real-world applications in considering images taken under uncontrolled environments

This paper consists of seven sections. First section is the introduction section, which covers background study, motivation and incitement, and contribution. Section 2 discusses the related work that has been accomplished in the area. The proposed FER model and the algorithms used to develop the model will be explained in Section 3. Then, section 4 describes the FER model's implementation in the experiment environment. Section 5 elaborates the results and discussion of the experiments, while section 6 concludes the paper. Finally, section 7 discuss the limitation and future works.

2. RELATED WORKS

According to [18], there are two primary combination approaches: i) prior-combination and ii) post-combination. Prior-combination is a technique for combining a set of extracted features [18]. Both approaches will result in a new set of features that is distinct from the initial sets. Several studies focused on fusing different features representation, such as the fusion of the whole face region and key expression regions. For instance, Jun *et al.* [19] developed a FER model based on the fusion of LBP features of local key expressions with global features. LBP features are used to extract local features in the face region, including eyes, eyebrows, between-eyebrow, nose, and mouth. Then, these features are fused into the features of the whole face (global features) and formed a new feature [19]. This method was able to preserve the overall features of the facial images. Shengtao *et al.* [7] proposed another fusion of the global and local feature with CNN for the FER problem. Using the FER-2013 dataset, the recognition rate for AlexNet, VGGNet, and ResNet is 66.6%, 69.41%, and 70.74%, respectively [7]. In [20], a method was proposed to improve the FER system's recognition rate by combining the entire face image with multiple sub-regions. The recognition rate reported using the proposed method are 99.07%, 95.95%, 67.7%, and 59.97% on Cohn-Kanade (CK+), Japanese female facial expression (JAFFE), FER-2013, and static facial expressions in the wild (SFEW) datasets, respectively.

The post-combination process either improves classification performance with the assistance of a second classifier or collects multiple results in order to vote on the most frequently occurring result to be selected as the final outcome. Several papers, such as [12], [21], and [22], employ an ensemble of classifiers to determine the sentiment of an image. Liu *et al.* [12] improved the FER rate by developing a method based on the CNN ensemble. The face image is initially fed into each of the three CNN subnets and trained separately. Concatenation of the extracted features occurs when a fully connected layer is inserted at the end of the subnets. Based on the evaluation using the FER-2013 dataset, the proposed FER model in [12] obtained an overall accuracy of 65.03%. Huang *et al.* [22] developed a framework for multimodal expression recognition that integrates facial expression and electroencephalography (EEG) data by utilizing a two

decision-level fusion method based on the enumerate weight rule and adaptive boosting techniques. Liu *et al.* [23] proposed a merged convolutional neural network (MCNN) approach for enhancing the robustness and accuracy of real-time FER. The MCNN architecture is composed of the partial ResNet and improved LeNet architectures, and it concatenates the feature maps of facial expressions extracted using these architectures. The work by Jia *et al.* [24] employed ensemble learning to integrate the output of three CNNs: i) AlexNet, ii) VGGNet, and iii) ResNet, using an SVM classifier. On the FER-2013 dataset, the ensemble-based FER model achieved a 71.27% accuracy. Based on the recent studies on FER, an ensemble model can outperform any single contributing model in terms of prediction and performance. This paper proposed an ensemble-based FER model based on three CNN models.

3. PROPOSED METHOD

This article proposed an ensemble-based approach for classifying facial expression into seven categories: two positive representations (happy and surprise), four negative representations (angry, disgust, fear, and sadness), and neutral facial expression. Ensemble learning's central concept is to train multiple base learners as ensemble members and then combine their predictions into a single output that can potentially outperform any other ensemble member with uncorrelated error on target data sets [25]. In a real-time application, sentiment analysis required faster and accurate recognition. Models that are pre-trained on ImageNet like Resnet50 and InceptionV3 are good at detecting high-level features like patterns, edges, and many more, which helps in faster convergence. This paper used the pre-trained models with transfer learning (TL) by fine-tuning the last predicting layers of the model to make them more relevant for FER. The main advantage of using TL is that it reduces training time and generalization error.

Figure 1 depicts the block diagram of the proposed FER model. The process begins with pre-processing and continues with classification using three different CNN architectures. Pre-processing is a process used to maximize the functionality of the FER model. CNN is a deep learning algorithm that takes input images, assign importance (learnable weights and biases) to various aspects or objects in the image, and differentiate one from the other [26]. CNN is generally made up of three types of layers: convolution, pooling, and fully connected layers [27]. Convolution and pooling layers perform feature extractions, whereas the fully connected layer maps the extracted features into the final output (classification) [27]. For this paper, the proposed FER model utilized three different CNN architectures; i) custom CNN ii) ResNet50, and iii) InceptionV3. After the pre-processing phase, the input is then fed into the custom CNN, ResNet50, and the InceptionV3 model in parallel. The architecture of these models will be further elaborated in subsections 3.1, 3.2, and 3.3, respectively. Finally, the final expression classification is determined by averaging the weight values output by the three individual classification models through the model averaging method.

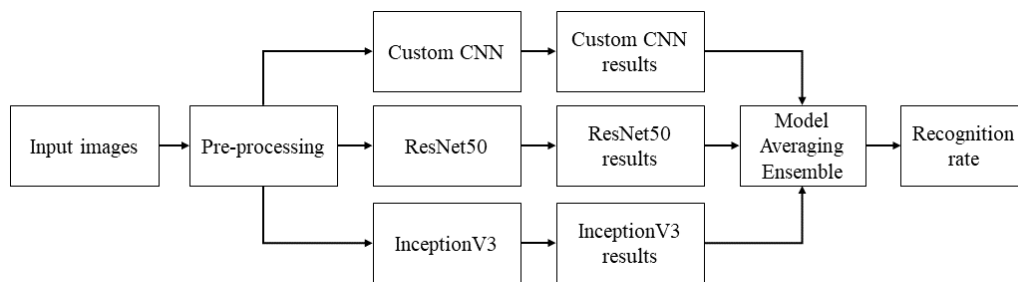


Figure 1. Block diagram of the proposed FER model

3.1. Custom CNN

The custom CNN model is a classical inspired by the work in [28], which gives promising results as a member of different ensemble models. This model is made up of four convolutional blocks, one fully connected layer, and one Softmax classification layer. It receives an image with a resolution of 64-by-64 as input. The first block comprises two convolution layers, one BatchNorm layer, one maximum pooling layer, and one dropout layer. Then, the remaining three blocks are composed of the same layers as the first, plus two BatchNorm layers. The kernel size chosen for the convolution layers is three-by-three dimensions, which is considered the smallest kernel size to capture the surrounding information. Following that, there are four layers in the fully connected layer, with the last layer serving as the output layer and employing Softmax as

the loss function. The final layer is made up of seven nodes that correspond to seven expressions. The probability of each expression is computed, and then the node with the largest probability value is chosen as the output class.

3.2. ResNet50 (residual network)

ResNet50 is a variant of the ResNet model, which has 50 layers, which consist of 48 convolutional layers, one max pool layer, and one average pooling layer [29]. It takes an image size of 197-by-197 as input and is fed into the network. The final layer of the last convolutional blocks has an output size of three-by-three-by-2048. The original classifier of ResNet50 is removed, and a new classifier that fits the problem addressed in this paper is added. At the end of the ResNet-50 model, a fully connected layer is applied, followed by an output layer. The fully connected layer contains 1024 nodes, while the final layer contains seven nodes representing the seven expression classes [29].

3.3. InceptionV3

InceptionV3 model was first introduced by Szegedy *et al.* [14]. InceptionV3 consists of six convolution layers, two pooling layers, three traditional inception modules, five factorized inception modules, and two reduced inception modules [14]. The patch size for the convolution layers is three by three dimensions. In this paper, the InceptionV3 received an input size of 139-by-139. Similar to ResNet50, the original fully connected layer and Softmax layer are removed. A global average pooling layer is added after the final convolution layer, followed by a dense layer with 1024 nodes and a seven-node output layer. Each corresponds to one of the seven facial expressions. The final layer of the last convolutional blocks in InceptionV3 networks gives output with three-by-three-by-2048 in dimension and yields approximately 18.8 million weights. By utilizing global average pooling, the weights of the model's trainable parameters can be reduced from 18.8 million to approximately 2.09 million.

3.4. Model averaging ensemble

The proposed FER model used the model averaging ensemble approach to combine the classification results from the three models (custom CNN, ResNet50, and InceptionV3) to perform the final classification. Figure 2 presented the pseudocode of the Model Averaging Ensemble algorithm. Each model will generate a probability vector assigned to each of the class labels. The three models' predictions are combined by summing the probabilities for each class prediction and returning the prediction index with the highest probability value via the NumPy argmax function. The label in the index generated by the NumPy argmax function will be the new classification result of the model averaging ensemble.

Algorithm 1: Model Averaging Ensemble

Input: P^k , Probability vector assigned by a classifier K to each class label
 C , class labels present in the dataset
 L , number of classifiers ($L = 3$)

Output: Prediction result R

1. Begin
2. for each j in C
3.
$$P_R \leftarrow P_j^K(x) = \frac{1}{L} \sum_{i=1}^L P_i(x)$$
4. end for
5. Normalize
$$P_R \leftarrow \frac{P_R}{\max(P^K)}$$
6. Compute $R = \arg \max(P_R)$
7. Return prediction R
8. End

Figure 2. Pseudocode of the classification using model averaging ensemble

4. EXPERIMENTAL SETUP

4.1. Experiment environment

The experiment is conducted using Keras and TensorFlow framework. The experiment environment is as follows: CPU clock speed is 2.60 GHz; 12 G of RAM, the graphics processing unit is NVIDIA GeForce GTX 950M, which has 4G of video memory. The operating system is Window 8.1 64-bit system.

4.2. Database selection

The database used in training and testing the proposed FER model is the FER-2013 dataset. It is an open-source dataset downloaded from Kaggle and was created by Piere-Luc Carrier and Aaron Courville. The FER-2013 dataset consists of 35887, 48 by 48 sized grayscale face images with seven expressions: angry (4953), disgust (543), fear (5121), happy (8989), sad (6077), surprise (4022), and neutral (6198) [30]. Table 1 summarized the dataset's partition for training, validation, and testing. The training dataset contains 28709 images, which accounts for approximately 80% of the total FER-2013 dataset. Another 20% of the dataset is divided equally between validations and testing. There are 3589 images in each of the validation and testing datasets. Table 2 presented the distribution of expression labels in the training, validation, and testing datasets.

Table 1. Dataset partition

Dataset	Total	Percentage (%)
Training	28709	80%
Validation	3589	10%
Testing	3589	10%

Table 2. Distribution of the expression labels in training, validation, and testing datasets

Label	Emotion	Total images per expression in Training dataset	Total images per expression in Validation dataset	Total images per expression in Testing dataset
0	Angry	3995	467	491
1	Disgust	436	496	528
2	Fear	4097	653	594
3	Happy	7215	607	626
4	Sad	4830	56	55
5	Surprise	3171	895	879
6	Neutral	4965	415	416

4.3. Data pre-processing

All the training, validation, and testing dataset are pre-processed to encode a batch of images. It yields a NumPy array containing the batch's shape and size, image height, image width, and several channels. All three models have undergone different rescaling processes because the required input image size of each model differs. The size of input for custom CNN, ResNet50, and InceptionV3 are 64-by-64, 197-by-197, and 139-by-139, respectively. Since the ResNet50 and InceptionV3 models receive three input channels, the grayscale image with only one channel is expanded into three channels by copying the grayscale information to the other two channels to meet the image ResNet50 and InceptionV3 model's format requirements. Next, zero-mean normalization is performed to standardize the inputs to the next subsequent layer for each mini-batch to stabilize the learning process. Figure 3 summarizes the data pre-processing on the input data of the three neural network models.

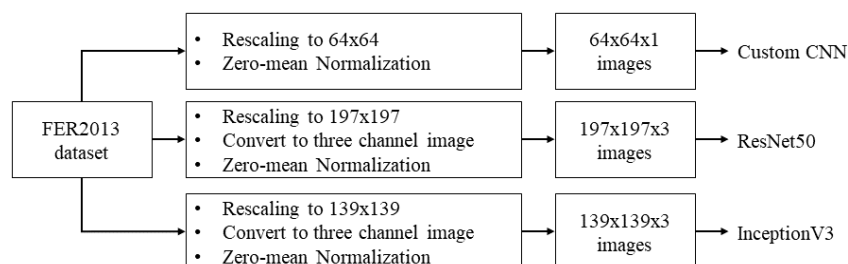


Figure 3. Data pre-processing process

4.4. Parameters initialization

Each model will undergo parameter updating during the training phase to find the best fit for the model. The initialized parameters included: type of optimizer, learning rate, batch size, number of epochs, momentum, and weight decay values (Beta1 and Beta2). The batch size determines how many samples are fed into the model. The epochs specify how many times the algorithm should be trained on the entire dataset.

The parameter for the custom CNN model was initialized as follows: the Adam optimizer with a learning rate of 0.001 was used. The batch size had been set to 64. The epochs can only be less than or equal to 100. Beta1 and Beta2 were set to 0.9 and 0.999, respectively. Finally, to avoid division by zero in the implementation, the epsilon was set to $1e^{-7}$.

The initialization of the parameters in the ResNet50 and InceptionV3 models is similar. The parameters for these two pre-trained models were set up as follows: The stochastic gradient descent (SGD) optimizer was used, with a learning rate of 0.0001. Same as the custom CNN model, the batch sizes were set to 128, and the epochs were set to be less than or equal to 100. Furthermore, the momentum was set to 0.9, while the weight decay was set to zero. Finally, the Nesterov momentum was set to true.

Additionally, a strategy of early stopping was used to avoid overfitting. When the loss on the validation set no longer drops after the specified number of epochs (patience mode), the training is terminated early. All three models are subjected to the strategy. The number of epochs varies between 20-100.

4.5. Training phase

Figure 4 illustrates the process involved during the training phase. In this phase, the three single classifiers: i) custom CNN, ii) ResNet50, and iii) InceptionV3 are separately trained on the training dataset while fine-tuning the optimal parameters using the validation dataset. First, the processed training data is fed into the single classifier and trained layer by layer to learn the images' features. Next, the features extracted from the model are fed into the fully connected layer, which is then used to classify the seven expressions. During the training phase, the validation dataset is used to select the optimal parameter set and saves the optimal model. Selecting the optimal parameters is repeated until there is no reduction in the validation loss or the process automatically stops when the number of epochs exceeds 100.

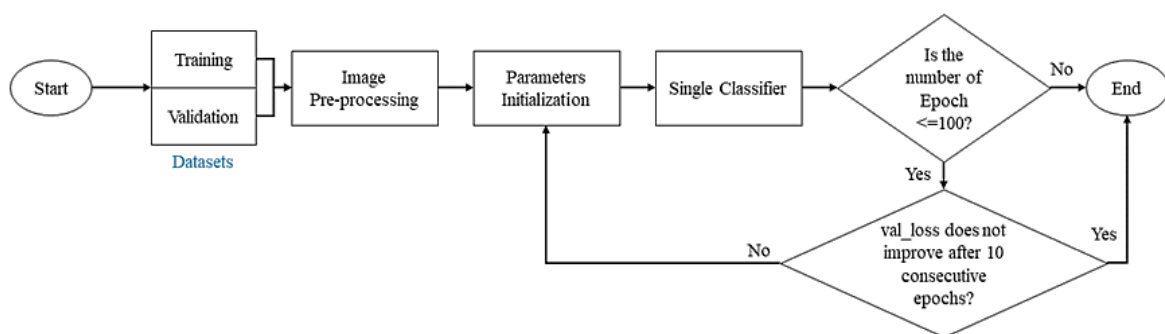


Figure 4. Flowchart of the training phase process

4.6. Testing phase

The purpose of the testing phase is to evaluate the performance of the classification models based on the testing dataset. Figure 5(a) illustrates the testing phase for a single classifier. Whereas Figure 5(b) illustrates the testing phase for the ensemble-based approach, which implemented the model averaging method for the final prediction of the facial expression. Model averaging is a technique for ensemble learning in which each member of the ensemble contributes equally to the final prediction, and the prediction of each member are completely uncorrelated. Specifically, each trained model predicts seven class labels; disgust, anger, fear, sad, surprise, happy, and neutral. The final prediction can then be converted to a class label by invoking NumPy's argmax function on the predicted probabilities and returning the prediction index with the highest probability value as the final class label. Recognition accuracy with one single-expression feature is relatively low in many cases [15], [16], [31]. The face regions, such as the mouth, nose, and eyes, are sensitive organs that show the different significance that determine one facial expression. This study provides a classification based on the strengths of ensemble methods and the significance of sensitive component characteristics in facial expression recognition. To minimize error rates, the models must produce output predictions that are relatively uncorrelated [32].

4.7. Performance metric

The accuracy metric is used as the performance metric to measure the model's overall performance on the testing set, supposed that confusion matrix (CM) is a confusion matrix of n -by- n dimension, where n is

the total number of different facial expressions. Furthermore, the row of CM represents the actual expression, while the column of CM represents the predicted expression. Given that seven facial expressions are used as the class outputs in this study, the value of n is 7. Finally, let $C_{i,j}$ indicates the CM cell's value at index row i and column j , where $i, j = 1, 2, \dots, n$. The accuracy metric is defined as in (1).

$$accuracy = \frac{\sum_{i,j=1}^n C_{i,j}}{\sum_{i=1}^n \sum_{j=1}^n C_{i,j}} \quad (1)$$

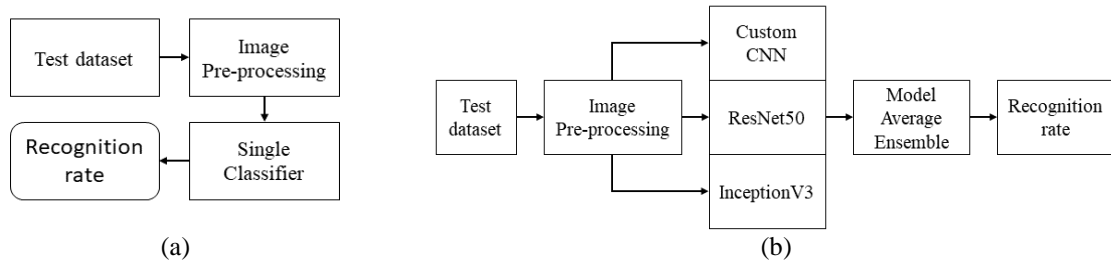


Figure 5. Flowchart of the testing phase for (a) single classifier and (b) model averaging ensemble classifier model

5. RESULTS AND DISCUSSION

5.1. Investigation of the effect of facial expression feature on the FER performance

A preliminary experiment is carried out using the Japanese female facial expression (JAFPE) dataset as a baseline. This experiment investigates the effect of facial expression features on FER performance and selects the best features to be used in the proposed FER model. The JAFPE dataset is relatively small in comparison to FER-2013. As a result, the time required to train the model on the JAFPE dataset is short, thus significantly reducing the processing time required to investigate the effect of facial expression features on FER performance compared to using the FER-2013 dataset. Based on the previous studies, several papers show pretty convincing results using the LBP feature [9], [22] and the CNN feature [7], [12]. Therefore, three experiments are undertaken in the preliminary stage: classification using LBP features, CNN features, and a fusion of LBP and CNN features (denoted as LBP+CNN). The robust local binary pattern (RLBP) [33] algorithm is used to extract the LBP-based features in the input images.

On the other hand, the CNN-based features are extracted using the convolutional base of the custom CNN model. For the fusion-based feature representation (LBP+CNN), the extracted LBP features and CNN features are concatenated as one features vector. Lastly, the classifier part of the custom CNN model is used in the classification phase using these feature representations. The whole process of the preliminary experiments conducted is summarized in Figure 6. The feature representation with the highest performance will then be used in all subsequent experiments. The training set comprises 80% of the JAFPE dataset, while the testing set comprises 20% of the JAFPE dataset. The data are shuffled before splitting, resulting in 170 images as the training set and 43 images as the testing set. Table 3 summarized the classification results using LBP, CNN, and LBP+CNN features as the feature representation on the JAFPE dataset. As shown in Table 3, the overall accuracy performance obtained using the LBP features, CNN features, and LBP+CNN features are 72.1%, 81.4%, and 79.1%, respectively. It was discovered that classification using CNN features has a higher recognition rate than classification using LBP features or LBP+CNN features. Thus, the CNN features will be used for all the subsequent experiments.

5.2. Experiment with a single classifier on the FER-2013 dataset

Custom CNN, Resnet50, and Inception V3 are trained and tested on the FER-2013 dataset. Table 4 summarized the classification results for custom CNN, ResNet50, and InceptionV3 models. On the other hand, Table 5 presented the performance of the single classifiers with average accuracy sorted from best to worst. The average accuracy is defined as in (2).

$$Average\ accuracy = \frac{Custom\ CNN\ accuracy + ResNet50\ accuracy + InceptionV3\ accuracy}{3} \quad (2)$$

For the custom CNN model, 278 out of 491 images are correctly classified as having an angry expression. Following that, 33 of 55 disgust expression images are correctly classified, while 294 of the 528

images for the fear expression are correctly classified. Next, for the happy expression, 740 of 879 are correctly classified, while for the sad expression, 286 out of 594 were correctly classified. Subsequently, surprise expression has 328 images correctly classified out of the 416 images. And lastly, for the neutral expression, 405 images out of 626 images are correctly predicted as neutral. In general, the custom CNN model performed poorly when classifying sad expressions but excelled when classifying happy expressions.

Based on Table 5, the ResNet50 model gives more than 60% accuracy values for all the expressions, except for fear and sad expression, 48.7%, and 59.9%, respectively. This model correctly classified 314 images from 491 anger expression images and 36 images from 55 disgust expression images as shown in Table 4. Following that, the ResNet50 model has correctly classified 799 out of 879 happy expression images, contributing to 90.9% accuracy. Furthermore, the model also gives 79.6% accuracy in classifying the surprise expression by correctly classifying a total of 331 out of 416 surprise expression images. Lastly, for the neutral expression, 464 images out of 626 images are correctly classified as neutral expression. ResNet50, like custom CNN, performs best when classifying happy expressions. While not identical to the custom CNN model, the ResNet50 model performed the worst when classifying negative expressions (fear expression). The InceptionV3 model correctly classified 271 images from 491 anger expression images. This model performs significantly worse than expected when classifying disgust and fear expressions, with 36.4% and 38.1% accuracy, respectively. On the other hand, the InceptionV3 model achieves a high accuracy of 88.6% when classifying 779 out of 879 happy expression images. Subsequently, InceptionV3 correctly classified 296 out of 594 sad expression images and 311 images out of 416 surprise expression images. Lastly, for the neutral expression, 414 images out of 626 images are correctly classified as neutral expression images. Compared to the custom CNN and ResNet50 models, the InceptionV3 model also performs worse at classifying negative expressions (disgust, fear, and sad) and better at classifying positive expressions (happy, surprise).

From Table 5, it can be seen that all models perform optimally when classifying positive expressions (happy and surprised), with happy expressions performing the best overall. In contrast, all the models perform poorly when classifying negative expressions (anger, disgust, sadness, and fear), with fear expression being the worst results overall. This implies that the inter and intra-expression differences between negative expressions in the FER-2013 database may not be significant.

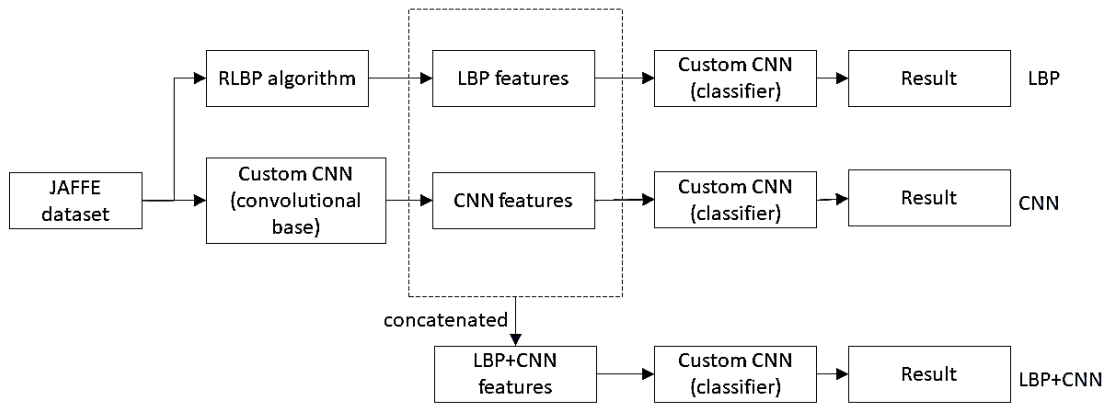


Figure 1. Summary of the preliminary experiments using three feature representations: BP features, CNN features, and LBP+CNN features

Table 3. Summary of accuracy performance using LBP, CNN, and LBP+CNN features as feature representations evaluated on JAFFE dataset

Expression	Accuracy for LBP features (%)	Accuracy for CNN features (%)	Accuracy for LBP+CNN features (%)
Angry	100	75	100
Disgust	75	62.5	75
Fear	28.5	85.7	57.1
Happy	71.4	100	85.7
Sad	75	75	100
Surprise	83.3	83.3	50
Neutral	85.7	85.7	100
Overall Performance	72.1	81.4	79.1

Table 4. Summary of classification results of Custom CNN, ResNet50, and InceptionV3 models on FER-2013 dataset

Expression	Total Images	Correctly classified (Custom CNN)	Misclassified (Custom CNN)	Correctly Classified (ResNet50)	Misclassified (ResNet50)	Correctly Classified (InceptionV3)	Misclassified (InceptionV3)
Angry	491	278	213	314	177	271	220
Disgust	55	33	22	36	19	20	35
Fear	528	294	234	257	271	201	327
Happy	879	740	139	799	80	779	100
Sad	594	286	308	356	238	296	298
Surprise	416	328	88	331	85	311	105
Neutral	626	405	221	464	162	414	212
Total	3589	2364	1225	2557	1032	2292	1297

Table 5. A comparison of the single classifiers' accuracy with average accuracy sorted from best to worst

Expression	Accuracy for Custom CNN (%)	Accuracy for ResNet50 (%)	Accuracy for InceptionV3 (%)	The average accuracy (%)
Happy	84.2	90.9	88.6	87.9
Surprise	78.8	79.6	74.8	77.73
Neutral	64.7	74.1	66.1	68.3
Angry	56.6	64	55.2	58.6
Disgust	60	65.5	36.4	53.97
Sad	48.1	59.9	49.8	52.6
Fear	55.7	48.7	38.1	47.5
Overall performance (%)	65.9	71.2	63.9	67.00

5.3. Experiment using the ensemble approach on the FER-2013 dataset

The confusion matrix for the proposed ensemble-based FER model is presented in Table 6. Based on Table 6, two significant findings are observed. First, the disgust and angry expressions interfered with each other easily, considering the 11 out of 55 disgust images misclassified as angry. This observation might be due to the early phase of dynamic facial expression between anger and disgust. As shown in Figure 7, the angry expression in Figure 7(a) and the disgust expression in Figure 7(b) is quite resembled due to the similar lip funneler and the aligned movement of the nose wrinkle. This is supported by [34], where it is reported that there are certain discrepancies in emotion, such as anger as a result of irritability. Secondly, the surprise and fear expressions are almost identical. There are 46 out of 528 fear images misclassified as a surprise, and 35 out of 416 surprise images misclassified as fear. By examining both surprise in Figure 7(c) and fear in Figure 7(d), it is possible to see how the jaw drops and the upper lip raises generate an almost identical image. In general, it was found that the proposed FER model performs well when it comes to classifying positive expressions (happy and surprise). In contrast, the proposed FER model performs poorly when it comes to classifying negative expressions (anger, disgust, fear, and sad), with the lowest accuracy for fear expressions. The accuracy of each class is compared in Table 7 between the custom CNN, ResNet50, InceptionV3, and the proposed model.

Based on Table 7, it can be seen that the proposed FER model outperforms all other models when it comes to classifying positive expressions (91.7% for happy and 81.7% for surprise) and neutral expression (76.5%). In contrast, the proposed FER model is not the best performing model in classifying negative expressions. Except for the fear expression, ResNet50 is the best model for classifying all negative expressions (anger, disgust, and sadness). One possible reason for this drawback could be the poor performance of the InceptionV3 model in detecting disgust expression. The misclassifications made by the InceptionV3 model as a member of the ensemble method have greatly affected the final prediction lower the classification performance. Thus, it can be concluded that the proposed FER model is the most effective at detecting positive emotions. However, when it comes to classifying disgust, anger, and sadness, the ResNet50 model is the best choice. Although the custom CNN has the best performance in classifying fear expression (55.7%), the proposed FER model can still classify fear expression with comparable performance (52.8%). These results suggest that combining multiple classification models can improve recognition rates by contradicting one another's misclassifications. InceptionV3 has the worst performance compared to the other individual models, with ResNet50 being the best performing individual model (71.2%). By ensembling the three models, the proposed FER models improve accuracy by 1.1% compared to the best performing individual model (ResNet50), resulting in an overall accuracy of 72.3%.

Table 8 compared the proposed FER model to some of the previous works on the FER-2013 dataset. Most of the ensemble-based approaches [7], [26], including the proposed FER model, give higher accuracy compared to the non-ensemble-based methods. In contrast to the approach used in this paper, the work by Liu

et al. [12], Wang *et al.* [20], and Jia *et al.* [24] employed the fusion of features extracted by different models, concatenated together using one classifier such as fully connected layers. Although the non-ensemble-based approach by Jha *et al.* [35] had better accuracy than ensemble-based methods by Wang *et al.* [20] and Liu *et al.* [12], the use of multiclass SVM loss function in their design results in a complex training system. Despite the fact that the approach by Gu *et al.* [7] has the advantage with the use of local face region in their design, the proposed FER model gives a comparatively higher accuracy on the FER-2013 dataset. In addition, the findings in Table 8 proved that the use of deeper networks like ResNet, AlexNet, and InceptionV3 could improve the FER performance.

Table 6. The confusion matrix of the proposed FER model tested on the FER-2013 dataset

Expression	Angry	Disgust	Fear	Happy	Sad	Surprise	Neutral	Total	Accuracy (%)
Angry	313	3	45	18	64	7	41	419	63.7
Disgust	11	33	7	0	2	1	1	55	60
Fear	58	3	279	10	80	46	52	528	52.8
Happy	11	0	8	806	13	15	26	879	91.7
Sad	45	1	64	21	344	4	115	594	57.9
Surprise	8	0	35	16	8	340	9	416	81.7
Neutral	15	0	24	25	77	6	479	626	76.5



Figure 7. Sample of FER-2013 dataset for (a) angry, (b) disgust, (c) surprise, and (d) fear

Table 7. A comparison of classification results with proposed FER model accuracy sorted from best to worst

Expression	Accuracy for custom CNN (%)	Accuracy for ResNet50 (%)	Accuracy for InceptionV3 (%)	Accuracy for the proposed FER model (%)
Happy	84.2	90.9	88.6	91.7
Surprise	78.8	79.6	74.8	81.7
Neutral	64.7	74.1	66.1	76.5
Anger	56.6	64	55.2	63.7
Disgust	60	65.5	36.4	60
Sad	48.1	59.9	49.8	57.9
Fear	55.7	48.7	38.1	52.8
Overall performance	65.9	71.2	63.9	72.3

Table 8. Comparison between the proposed FER model and previous studies using the FER-2013 dataset

Authors	Method's Significance in Feature selection and classification	Ensemble-based	Use pre-trained model	Accuracy on FER-2013 dataset
Our proposed method	Classification results from custom CNN, ResNet50, and InceptionV3 are ensemble using the Model Averaging method	Yes	Yes	72.3%
Jia <i>et al.</i> [24]	Feature extracted using AlexNet, VGGNet, and ResNet are fed into SVM for classification	Yes	Yes	71.27%
Shengtao <i>et al.</i> [7]	Ensembled two ResNet models re-trained using original and cropped data respectively.	Yes	Yes	70.74%
Jha <i>et al.</i> [35]	CNN with a multiclass SVM loss function	No	No	69.9%
Wang <i>et al.</i> [20]	The cropped sub-regions and the whole face images fed into four designed CNN models	Yes	No	67.7%
Liu <i>et al.</i> [12]	Features extracted using three different subnets (CNN) are concatenated using fully connected layers	Yes	No	65.03%
Jadhav <i>et al.</i> [36]	CNN	No	No	63%

6. CONCLUSION

Provide a statement that what is expected, as stated in the introduction section can This paper proposed CNN's ensemble-based FER model for classifying facial images into seven sentiments: happy, surprise, sad, disgust, anger, fear, and neutral. This paper also successfully achieved the two objectives: i) to develop a FER model using an ensemble of existing FER models and ii) to evaluate the proposed FER model based on the FER-2013 dataset. First, a structured CNN was designed, and two pre-trained models (ResNet50 and InceptionV3) were re-architected to act as the base model. The well-trained neural network models were then ensembled using the model averaging method to decide the final classification of the facial expression. The proposed FER model's performance was evaluated using the FER-2013 dataset, a dataset captured under an uncontrolled environment. By ensembling the models together, it has achieved 72.3% accuracy for facial expression recognition. The key advantage of the proposed FER model is that it emphasizes the usage of multiple CNN architectures rather than just one. It is possible to get better performance by ensemble all the results together because the members of the ensembled models contradict each other's misclassification. In addition, it was found that a certain classification model could give better performance in detecting a particular facial expression. Specifically, the ResNet50 model performs the best in identifying negative emotions like disgust, anger, and sadness. On the other hand, the proposed FER model is the best model for classifying positive emotions like happy, surprise, and neutral.

7. LIMITATION AND FUTURE WORKS

The proposed ensemble-based FER model shows low performance in detecting negative emotions compared to a single classifier. In the future, the proposed ensemble-based FER model could be improved further by incorporating a single classifier that excels at classifying negative emotions specifically for ensemble-based negative emotion FER. The second limitation concerns the angle of face detection. The expression recognition algorithm is trained using the frontal images, so bringing the proposed FER model into the real-world application might give a lower performance than the results presented in this paper. In order to train the model in recognizing expression at different angles, future researchers should expand the dataset to include expression at different angles. Last but not least, the effect of the underlying features extracted from each class's image input on the classes' individual performance and overall performance can be investigated in future research.

ACKNOWLEDGEMENTS

This research was funded by the Research Management Centre (PPP), Universiti Malaysia Sabah, under grant number GA19095.





REFERENCES

- [1] K. Chakraborty, S. Bhattacharyya, and R. Bag, "A survey of sentiment analysis from social media data," *IEEE Transactions on Computational Social Systems*, vol. 7, no. 2, pp. 450–464, Apr. 2020, doi: 10.1109/TCSS.2019.2956957.
- [2] E. G. Mounq, J. A. Dargham, A. Chekima, and S. Omatu, "Face recognition state-of-the-art, enablers, challenges and solutions: A review," *International Journal of Advanced Trends in Computer Science and Engineering*, vol. 9, no. 1 Special Issue 2, pp. 96–105, 2020, doi: 10.30534/ijatcse/2020/1691.22020.
- [3] J. A. Dargham, A. Chekima, and E. G. Mounq, "Fusing facial features for face recognition," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 1, no. 5, pp. 565–572, 2012, doi: 10.9781/ijimai.2012.157.
- [4] J. A. Dargham, A. Chekima, E. Mounq, and S. Omatu, "Data fusion for face recognition," in *Advances in Intelligent and Soft Computing*, 2010, pp. 681–688.
- [5] K. Patel *et al.*, "Facial sentiment analysis using AI techniques: State-of-the-art, taxonomies, and challenges," *IEEE Access*, vol. 8, pp. 90495–90519, 2020, doi: 10.1109/ACCESS.2020.2993803.
- [6] B. Islam, F. Mahmud, and A. Hossain, "Facial expression region segmentation based approach to emotion recognition using 2D gabor filter and multiclass support vector machine," *2018 21st International Conference of Computer and Information Technology, ICCIT 2018*, pp. 21–23, 2019.
- [7] G. Shengtao, X. Chao, and F. Bo, "Facial expression recognition based on global and local feature fusion with CNNs," in *2019 IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC)*, Sep. 2019, pp. 1–5, doi: 10.1109/ICSPCC46631.2019.8960765.
- [8] Q. Xu and N. Zhao, "A facial expression recognition algorithm based on CNN and LBP feature," in *2020 IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC)*, Jun. 2020, pp. 2304–2308, doi: 10.1109/ITNEC48623.2020.9084763.
- [9] M. Murtaza, M. Sharif, M. AbdullahYasmin, and T. Ahmad, "Facial expression detection using Six Facial Expressions Hexagon (SFEH) model," in *2019 IEEE 9th Annual Computing and Communication Workshop and Conference (CCWC)*, Jan. 2019, pp. 0190–0195, doi: 10.1109/CCWC.2019.8666602.
- [10] V. Perez-Gomez, H. V. Rios-Figueroa, E. J. Rechy-Ramirez, E. Mezura-Montes, and A. Marin-Hernandez, "Feature selection on 2D and 3D geometric features to improve facial expression recognition," *Sensors*, vol. 20, no. 17, pp. 4847–486, Aug. 2020, doi: 10.3390/s20174847.
- [11] T. Sinziana, T. Tudor, M. Cristian, and D. Laura, "An initial study of feature extraction's methods in facial expression




- recognition," *Proceedings - 2019 IEEE 15th International Conference on Intelligent Computer Communication and Processing, ICCP 2019*, pp. 251–257, 2019, doi: 10.1109/ICCP48234.2019.8959538.
- [12] K. Liu, M. Zhang, and Z. Pan, "Facial expression recognition with CNN ensemble," in *2016 International Conference on Cyberworlds (CW)*, Sep. 2016, pp. 163–166, doi: 10.1109/CW.2016.34.
- [13] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90.
- [14] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," *arxiv.org/abs/1512.00567*, Dec. 2015.
- [15] S. M. González-Lozoya, J. de la Calleja, L. Pellegrin, H. J. Escalante, M. A. Medina, and A. Benitez-Ruiz, "Recognition of facial expressions based on CNN features," *Multimedia Tools and Applications*, vol. 79, no. 19–20, pp. 13987–14007, May 2020, doi: 10.1007/s11042-020-08681-4.
- [16] Y. Pratama, L. M. Ginting, E. H. L. Nainggolan, and A. E. Rismanda, "Face recognition for presence system by using residual networks-50 architecture," *International Journal of Electrical and Computer Engineering*, vol. 11, no. 6, pp. 5488–5496, 2021.
- [17] K. Slimani, M. Kas, Y. El Merabet, Y. Ruichek, and R. Messoussi, "Local feature extraction based facial emotion recognition: a survey," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 10, no. 4, pp. 4080–4092, Aug. 2020, doi: 10.11591/ijece.v10i4.pp4080-4092.
- [18] F. A. Abd Almuhsen and Z. A. Khalaf, "Review of different combinations of facial expression recognition system," *Journal of Physics: Conference Series*, vol. 1591, no. 1, Jul. 2020, doi: 10.1088/1742-6596/1591/1/012020.
- [19] H. Jun, C. Jian-feng, F. Ling-zhi, and H. Zhong-wen, "A method of facial expression recognition based on LBP fusion of key expressions areas," in *The 27th Chinese Control and Decision Conference (2015 CCDC)*, May 2015, pp. 4200–4204, doi: 10.1109/CCDC.2015.7162668.
- [20] Y. Wang, Y. Li, Y. Song, and X. Rong, "Facial expression recognition based on auxiliary models," *Algorithms*, vol. 12, no. 11, pp. 227–242, Oct. 2019, doi: 10.3390/a12110227.
- [21] Y. Luo, L. Zhang, Y. Chen, and W. Jiang, "Facial expression recognition algorithm based on reverse co-salient regions (RCSR) features," in *2017 4th International Conference on Information Science and Control Engineering (ICISCE)*, Jul. 2017, pp. 326–329, doi: 10.1109/ICISCE.2017.75.
- [22] Y. Huang, J. Yang, S. Liu, and J. Pan, "Combining facial expressions and electroencephalography to enhance emotion recognition," *Future Internet*, vol. 11, no. 5, pp. 105–122, May 2019, doi: 10.3390/fi11050105.
- [23] K. C. Liu, C. C. Hsu, W. Y. Wang, and H. H. Chiang, "Facial expression recognition using merged convolution neural network," in *2019 IEEE 8th Global Conference on Consumer Electronics, GCCE 2019*, 2019, pp. 296–298.
- [24] C. Jia, C. L. Li, and Z. Ying, "Facial expression recognition based on the ensemble learning of CNNs," *ICSPCC 2020 - IEEE International Conference on Signal Processing, Communications and Computing, Proceedings*, pp. 0–4, 2020, doi: 10.1109/ICSPCC50002.2020.9259543.
- [25] Y. Yang, "Chapter 4 - Ensemble learning," in *Temporal Data Mining Via Unsupervised Ensemble Learning*, 2017, pp. 35–56.
- [26] S. Sumit, "A comprehensive guide to convolutional neural networks: the ELI5 Way," *Towards Data Science*.
- [27] R. Yamashita, M. Nishio, R. K. G. Do, and K. Togashi, "Convolutional neural networks: an overview and application in radiology," *Insights into Imaging*, vol. 9, no. 4, pp. 611–629, Aug. 2018, doi: 10.1007/s13244-018-0639-9.
- [28] B.-K. Kim, S.-Y. Dong, J. Roh, G. Kim, and S.-Y. Lee, "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)*, Jun. 2016, pp. 1499–1508, doi: 10.1109/CVPRW.2016.187.
- [29] R. I. Bendjillali, M. Beladgham, K. Merit, and A. Taleb-Almed, "Illumination-robust free face recognition based on deep convolutional neural networks architectures," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 18, no. 2, pp. 1015–1027, 2019.
- [30] I. J. Goodfellow *et al.*, "Challenges in representation learning: A report on three machine learning contests," *Neural Networks*, vol. 64, pp. 59–63, Apr. 2015, doi: 10.1016/j.neunet.2014.09.005.
- [31] A. Alreshidi and M. Ullah, "Facial emotion recognition using hybrid features," *Informatics*, vol. 7, no. 1, pp. 6–19, Feb. 2020, doi: 10.3390/informatics7010006.
- [32] S. Nagaraja and C. J. Prabhakar, "Extraction of curvelet based RLBP features for representation of facial expression," in *2014 International Conference on Contemporary Computing and Informatics (IC3I)*, Nov. 2014, pp. 845–850, doi: 10.1109/IC3I.2014.7019630.
- [33] D. Abbot, "Benefits of creating ensembles of classifiers," *TDAN.com*, 2001.
- [34] A. Freitas-Magalhaes, "Facial expression of emotion," *Encyclopedia of Human Behavior*, pp. 173–183, 2012.
- [35] V. Jha, P. D. Shenoy, and V. K. R., "Development of facial expression classifier using neural networks," in *2019 IEEE International WIE Conference on Electrical and Computer Engineering (WIECON-ECE)*, Nov. 2019, pp. 1–4, doi: 10.1109/WIECON-ECE48653.2019.9019937.
- [36] R. S. Jadhav and P. Ghadekar, "Content based facial emotion recognition model using machine learning algorithm," in *2018 International Conference on Advanced Computation and Telecommunication (ICACAT)*, Dec. 2018, pp. 1–5, doi: 10.1109/ICACAT.2018.8933790.

BIOGRAPHIES OF AUTHORS






Ervin Gubin Moug     is a senior lecturer in the Faculty of Computing and Informatics, Universiti Malaysia Sabah. His research interest generally falls under Computer Vision & Pattern Recognition, such as image processing, image segmentation, image classification, object detection, vision-based learning, and big data analytics. His domain of interest includes public health, smart health, agriculture, food security, biodiversity, and environmental sustainability. He received his Bachelor of Computer Engineering, Master of (Computer) Engineering, and Ph.D. in Computer Engineering from Universiti Malaysia Sabah (UMS) in 2008, 2013, and 2018, respectively. He can be contacted at email: ervin@ums.edu.my.






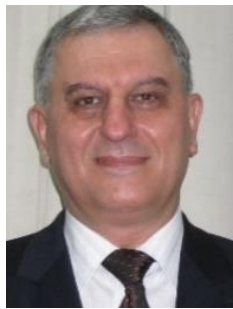
Chai Chuan Wooi    received his Bachelor of Network Engineering (Computer Science) from Universiti Malaysia Sabah (UMS) in 2019 and completed his Master's degree in Computer Science majoring in Computational Intelligence in January 2021. He continued work as a Research Assistant at Universiti Malaysia Sabah (UMS) in March 2021. His research focuses on developing face detection and face recognition for smart attendance purposes. He can be contacted at email: waynechai1995@gmail.com.






Maisarah Mohd Sufian    received her Bachelor of Electronic Engineering (Computer) from Universiti Malaysia Sabah (UMS) in 2020. She has been working as a Research Assistant at UMS since April 2021 and will officially begin her Master's degree in Computer Science in September 2021. Her research focuses on formulating a moment-invariant-based algorithm for COVID-19 detection using computerized tomography images. She can be contacted at email: maisarahmohdsufian@gmail.com.



Chin Kim On    is currently an Associate Professor in the Faculty of Computing and Informatics (Software Engineering Program) at Universiti Malaysia Sabah. Evolutionary computing, artificial neural networks, image processing, the Internet of Things (IoT), and biometric security systems are among his research interests. He is the author or co-author of 110 journal articles, book chapters, and conference proceedings. He is an IEEE Senior Member, a certified Computational Thinking Master Trainer, and a certified Professional Technologist. He can be contacted at email: kimonchin@ums.edu.my.



Jamal Ahmad Dargham    received his BSc in Control Systems Engineering from University of Technology, Iraq in 1984 and his MSc in Control Systems Engineering from the University of Manchester, UK, in 1987 and his PhD in Image Processing from Universiti Malaysia Sabah in 2008. From 1989 till 1996 he was a lecturer at the Advanced Management College, Sabah. From 1996 till now he has been with the Faculty of Engineering, University Malaysia Sabah. He was the Program Head of Computer Engineering Program from 2006 till 2011 and Head of the Artificial Intelligence Research Unit from 2016 to 2018. His main research interests are in Image Processing specifically Biometrics as well as Engineering Education. He has published more than 75 papers in refereed journals, conferences, book chapters and research reports. He has supervised more than 10 and was the examiner or more than 10 graduates at the Master and PhD levels. He can be contacted at email: jamalad@ums.edu.my.