Artificial Intelligence Applications And Innovations Vol. 1 No. 2 (2024)

Article history:
Received 5 Decembre 2023
Revised 5 February 2024
Accepted 17 February 2024
Published online 02 April 2024

Exploring A Novel Multi-Channel Structure to Improve Facial Expression Recognition On Occluded Samples Using Deep Convolutional Neural Network

Mohammad Hossein. Zolfagharnasab[1], Mohammad. Bahrani[1*], Masood. Hamed Saghayan[1], Fatemeh Sadat Masoumi[1]

[1] Department of Computer Science, Faculty of Statistics, Mathematics, and Computer, Allameh Tabataba'i University, Tehran, Iran

*** Corresponding author email address**: *bahrani@atu.ac.ir*

A r t i c l e I n f o

Article type:
Original Research

How to cite this article:
Zolfagharnasab, M. H., Bahrani, M., Hamed Saghayan, M., & Masoumi, F. S. (2024). Exploring a Novel Multi-Channel Structure to Improve Facial Expression Recognition on Occluded Samples Using Deep Convolutional Neural Network. Artificial Intelligence Applications and Innovations, *1*(2), 26-41 https://doi.org/10.61838/jaiai.1.2.3

A B S T R A C T

The development of Artificial Intelligence (AI) models with an accurate prediction of human facial expression has become a significant challenge for the cases in which masks and sunglasses cover critical facial areas. Given that a substantial portion of human interactions involves non-verbal communication, accurately detecting human emotions such as anger, fear, disgust, happiness, sadness, and surprise would benefit a wide range of applications, from security assessments to psychological treatments. As a workaround, the current study explores the performance of a novel multi-channel arrangement comprised of a Haar-wavelet, Histogram of Oriented Gradients (HOG), and grayscale filters to improve the predictions of deep Convolutional Neural Network (CNN) on occluded results. This study uses the FER-2013 dataset and produces occluded samples by applying a virtual mask that covers almost 55% of facial areas comprising the mouth, lips, and jaw locations. Further investigations, including the impact of each filter, utilizing pre-trained models on occluded samples (transfer learning), and comparison to prior models are also carried out. The proposed approach yields an accuracy rate of 71% for non-occluded and 66% for the occluded samples, which are 6% to 11% higher than the base model. Further transfer learning technique increases the accuracy metrics by 18%, indicating that non-occluded pre-trained models can reveal a broader range of features and their relation, which to some extent compensates for the removed features due to the occlusion. These results suggest the potential capabilities of the proposed technique for similar imaging applications.

Keywords: *Facial Expression Recognition, Partial Occlusion, Convolutional Neural Network, Transfer Learning, Histogram of Oriented Gradients, Haar Wavelet.*

1. Introduction

Expert psychologists indicate that more than half of our communication is done in a non-verbal context [1]. Such statistics can easily demonstrate that facial expressions play a significant role in daily communication [2]. In this regard, an area of research called Facial Expression Recognition (FER) has become a compelling subject for a wide range of studies from behavioral psychology to computer science [3]. As for computer science, FER is mainly followed in research such as facial identification [4], emotion recognition [5], and security systems [6]. The importance of facial detection was even more highlighted when the COVID-19 pandemic was officially announced, and countries made wearing masks an obligation [7]. Due to the noted compulsory laws, unpredicted challenges for facial

detection systems were produced, such as unlocking devices and systems responsible for the security and behavioral monitoring [8]. Therefore, the necessity of providing models capable of FER on occluded samples was felt [9].

The current study responds to the discussed obstacle by presenting a CNN model that accurately classifies occluded facial expressions. Although models based on the CNN architecture can conveniently detect significant attributes, the feature extraction step consumes considerable time [10]. Meanwhile, issues such as layer setup, hyperparameter tuning, and achieving a well-trained model increase as the network complexity grow [11].

The current study presents several suggestions to respond to such difficulties. First, instead of standard image formats such as RGB and grayscale, this study suggests constructing a novel multi-channel structure to highlight the influential features of an image sample. Here, two well-established filters in image-processing applications, which are HOG and Haar wavelet, are considered to highlight the significant features [12]. In fact, the multi-channel structure is expected to facilitate the feature extraction step since the input samples have inherited some degree of preprocessing, thus requiring less complex networks, fewer epoch counts, and lower training samples [13]. While such a technique is deployed for the FER applications, there are no limitations to extending such a strategy to other imaging issues [14]. Second, the current study has investigated the impact of transfer learning on developing models for occluded samples based on pre-existing models built with non-occluded datasets. The noted issue is vital because many pre-trained models for non-occluded FER are already available to improve the accuracy of occluded models if such a strategy is fruitful [15]. Last but not least, discrete assessments of each filter's performance on the classification quality and further comparison with prior studies are conducted. It must be noted that while the current study achieved admissible results using the described multi-channel structure coupled with VGG16, combining other multi-channel formats and classifiers can bring more promising results.

The remainder of this paper is presented in six more sections. In *Section 2*, the related works to the current study are reviewed. *Section 3* investigates dataset properties, applying filters and classifier properties. Next, details regarding the experiments, evaluation techniques, and limitations are provided in *Section 4*. The result analyses and further evaluation of the proposed model are included in *Section 5*. Next, recommendations regarding future works are presented in *Section 6*. Lastly, the conclusions are summarized in *Section 7*.

2. Related Work

Over the past decade, FER-involved studies have been growing due to the wide range of applications [16]. Boosted by COVID-19, the occluded FER studies have also gained much attention among researchers working on computer vision topics [17]. FER studies generally comprised three main steps: detecting facial areas, applying feature extraction, and classifying the images as one of seven general facial expressions: angry, Disgust, Fear, Happy, Sad, Surprised, and Neutral [18].

As the literature indicates, many studies involving neural networks use a variation of CNN architecture. The reason CNN models are more appealing to computer vision experts can be summarized into three key points. First, CNN models remove the necessity of manual feature extraction, which automates a significant portion preprocessing pipeline [19]. Second, CNN architectures are flexible in applying changes and can predict high-accurate results if sufficient samples are exposed during the training stage [20]. Third, CNN-based models have been used for a wide range of problems; therefore, many pre-trained models are available to accelerate the training process [21]. However, applying such techniques requires careful assessment to remove any possible mispredictions caused by the base model [22].

Unfortunately, though developments in the architecture of CNN models have automated feature extraction and classification steps; however, the training process has become significantly time-consuming when dealing with large datasets and high-quality samples [23]. In this regard, various modifications have been proposed in the literature to improve the time consumption of the feature extraction and the classification step while upholding the overall accuracy.

Among the techniques utilized to improve FER classifiers, many studies have employed hybrid classifiers to improve the prediction quality of the models [24]. For instance, a novel Searched Based Neural Network (SBNN) model, which was tuned to highlight facial features, was proposed by [25]. Based on their reports, an accuracy value of 70.02% was achieved from their model in the FER-2013 dataset.

A more detailed investigation is carried out by [26] in which a combination of a spatial transformer network (STN) and an Attentional Convolutional Neural Networks (ACNN) mechanism is proposed to improve the performance of the

ResNet-18 model. As the results indicate, their developed model has achieved an accuracy of 72.30% on the FER-2013 dataset.

A similar approach was followed in [26]; however, a Multi-Level Convolutional Neural Network (MLCNN) model was used as the classifier to highlight facial features. Their model has also achieved an accuracy of 73.03% on the FER-2013 dataset. Due to the success of the described approach, studies such as [27] developed more complicated hybrid models based on EfficientNet-Lite. After adequately optimizing parameters and hyperparameters with the k-value, their proposed model achieved an accuracy of 75.26% higher than prior models.

Another successful hybrid classifier was proposed by [28] in which a CNN model is combined with Scale Invariant Feature Transform (SIFT). The noted CNN-SIFT model has achieved an accuracy of 73.4% in the FER-2013 dataset. By applying a similar technique, models such as SOTA-EBNN utilized in [29] have combined three VGGs, two pre-trained models, and SVM and achieved an accuracy of 75.42% on the FER-2013 dataset. By Combining eight individual FER datasets such as FER-2013, CK, CK+, JAFFE, Chicago Face, FEI face, IMFDB, and TFEID, studies such as [30] have achieved an average accuracy of 74%, which is arguably accurate considering a gradual depth-wise and separable convolution was applied for down-sampling the original data.

Unlike the reviewed studies in which the performance of the CNN classifier was the primary target for the optimization [31], many studies have investigated techniques such as transfer learning, utilizing feature extracting filters, and data augmentation to improve the performance of FER applications. For instance, Pramerdorfer et al. [32] reviewed state-of-the-art FER studies conducted by CNN. Their study aimed to identify critical differences between the selected works and how each parameter improves/reduces the overall consistency of the model. Based on their conclusions, modern CNN architectures outperformed their older procedures if an appropriate arrangement of convolutional layers, pooling techniques, and dropout layers were considered. The impact of data transfer learning versus data augmentation is also investigated in several studies, such as [33]. Based on their results, data augmentation was a more effective technique than transfer learning to improve the classification quality in FER applications; however, the sample augmentation process was significantly more time-consuming and cumbersome in a wide range of studies. In conclusion, they have achieved an accuracy rate of 75.8% through the simultaneous use of both methods.

Inspired by classic Machine Learning (ML) techniques, some studies have also applied manual filters to highlight significant features before going through the training process of general FER applications [34]. For instance, the impact of the handcrafted HOG filter on the classification performance of CNN-based models is investigated in [35]. Their experiment was conducted on the FER-2013 dataset and achieved an accuracy of 72.1%. Similarly, studies such as [36] compared the performance of three popular feature extraction techniques, Gabor Filters (GF), Local Binary Pattern Operator (LBP), and Local Gabor Binary Pattern (LGBP), on the performance of the famous k-Nearest Neighbor (kNN) classifier on JAFFE dataset. Their results have also shown that applying the LGBP feature can improve the accuracy rate.

Based on the reviewed literature, it can be concluded that constructing multi-channel filters to improve the performance of deep neural networks has not been studied in previous studies. As a resolution, the current study evaluates the potential of utilizing multi-channel HOG, Haar, and grayscale filters to improve the performance of CNN models. In this regard, individual analysis is accomplished over the filters to evaluate their performance in highlighting the impactful features. Next, the simultaneous impact of the noted filters is carefully observed, and the results are compared with prior studies. Lastly, the effect of performing a non-occluded pre-training before developing the primary model is investigated. Such a transfer learning strategy is essential because many models are already trained with rich non-occluded samples. Such pre-trained models could serve occluded samples significantly well if the described approach is fruitful.

3. Methods

As described in prior sections, the current research comprises steps such as preprocessing the input data, removing unnecessary pixels (or noise), applying facial masks, and constructing multi-channel samples from the HOG, Haar wavelet, and grayscale filter output [37].

After the raw samples are processed, the training session of the CNN model can be initiated, and finally, the classification metrics are evaluated during the post-process session. Further details regarding each phase are described in the following sections.

3.1. Dataset

The FER-2013 is a dataset provided by the International Conference on Machine Learning (ICML) and introduced in a challenge on the Kaggle website [37]. The dataset consists of a training set with 28,709 instances and 3,589 for a test set. Fig. (1) illustrates several samples in the dataset alongside their categorization. Each sample in the FER-2013 dataset consists of a 48×48-pixel grayscale image of a human face. The distribution of FER-2013 samples among the classes is also presented in Table 1. Among the datasets corresponding to FER applications, FER-2013 is one of the most challenging datasets for classification since the sample resolution and the face angles are inconsistent throughout all images. In a study by [38], human accuracy on FER-2013 was close to 68%, even lower than state-of-art CNN classifiers. As indicated by Table 1, the samples corresponding to the Disgust class contain less than 2% of the dataset. Therefore, they are more likely to be misclassified due to the significant inhomogeneous distribution, as the reviewed literature suggests.

Figure 1

FER-2013 dataset

Table 1

Distribution of FER-2013 samples

Expression	Number of Samples	% of distribution
Angry	4953	13.82%
Disgust	547	1.52%
Happy	8989	25.11%
Sad	6077	16.98%
Fear	5121	14.22%
Surprise	4002	11.10%
Neutral	6198	17.25%

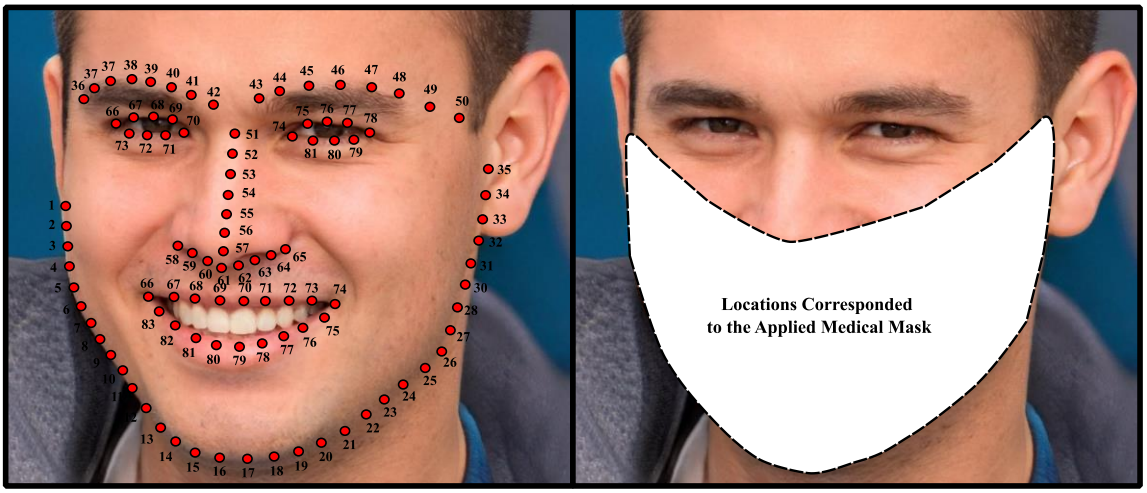
3.2. Occluded Sample Generation

Producing occluded sample from the FER-2013 dataset require additional steps. In this regard, the Dlib library is used to detect areas corresponding to significant facial elements such as eyes, eyebrows, jaw, and nose [39]. As depicted in Fig. (2a), facial features are identified by a sequence of numbers specified at each location. Next, the artificial medical mask is generated by connecting the red points so that the mask fully covers the areas from the jaws up to the middle nose area.

A schematic of the applied mask is depicted in Fig. (2b). Using the described technique, the same pair of images is available for both occluded and non-occluded images on the FER-2013 samples. As an important note, it should be stated that the area corresponding to the projected mask in this study is significantly wider than many prior studies focused

on occluded samples. For instance, in a survey by [35], only a single facial feature, such as the eyes, mouth, and nose, is covered during the occlusion, which is comparably less than applying a full medical mask that covers all the lower facial areas.

As a result, detecting facial expressions such as happiness in which the lower facial elements such as mouth, lips, and jaws are more impactful are more likely to be a challenge in the current study.



(a) Locations corresponded to facial features (b) Locations corresponded to the artificial mask

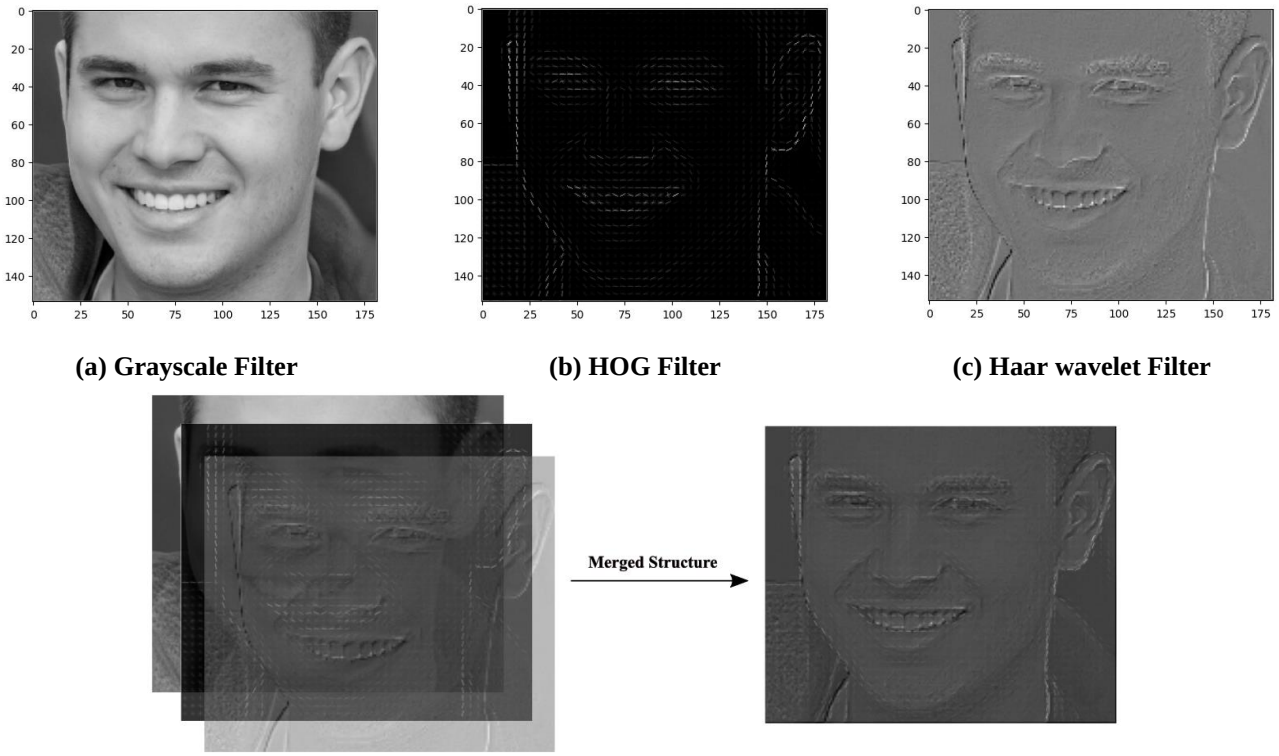
Figure 2

Generating occluded samples using Dlib

3.3. *Filter Properties*

Constructing the multi-channel structure is the second phase of the current preprocessing pipeline, which focuses on highlighting features of interest from a raw sample [40]. In conventional image-processing techniques, the feature extraction stage is carried out by sequentially applying proper filters to highlight the feature of interest. Such manual tasks are now conducted in deep neural networks; however, such training strategy is significantly time-consuming.

As a workaround, the current study selects two famous HOG and Haar filters to highlight significant features before the samples undergo training. As depicted in Fig. (3), the outputs created from these filters construct the multi-channel structure used as the CNN input. Although the current study selects HOG and Haar wavelet as the feature exposures, there are no limitations on using other filters.



(a) Grayscale Filter **(b) HOG Filter** **(c) Haar wavelet Filter**

(d) Merged Structure of Filters

Figure 3

Multi-channel images

3.3.1. HOG Filter

Targeted feature representation plays a vital role in the success of the FER models since it discriminates crucial information from the unnecessary portion. Among the popular filters that highlight effective pixels, the HOG filter is found to be an appropriate feature descriptor [41].

HOG extracts feature such as curved shapes formed on the eye, mouth, and jaw regions by calculating the pixelation gradients. The key point regarding HOG is the fact that it also highlights the pixel orientations, which is significantly beneficial when dealing with human expressions such as surprise and happiness in which gradients adjacent to the mouth, as well as regions closer to eyebrows, are essential for the classifier to detect FER [42].

3.3.2. Haar Wavelet Transform Response Filter

The Second filter selected in this study is the Haar wavelet transform response, a mathematical operation carried out on a sequence of rescaled and shaped small frequent Haar waves [43].

In short, the Haar filter extracts visual appearance from prominent facial regions at two different scale frequencies: black and white. Next, the Haar filter converts the image into the grayscale format and divides pixels as dark and bright from a certain threshold. By doing so, edges representing a feature of interest can be extracted by iterating over adjacent pixels of a particular area.

3.4. Classifier Properties

Selecting an appropriate classifier is essential for AI solutions since it is the main element responsible for identifying facial expressions [44]. Since there are no specific guidelines on the architecture of hidden layers for deep neural networks, AI frameworks are designed based on heuristic knowledge, iterative optimization, prior well-established models, and careful tuning of the involved hyperparameters [45]. Similarly, this study selects the conventional VGG16 framework as the base model since it is one of the most adopted network models in imaging applications [46]. Although other well-established models, such as GoogleNet and ResNet, were even more accurate,

the current study selects VGG16 since it involves lesser parameters when training from scratch. Moreover, one of the main objectives of the present study is to reduce the training process of DL models by utilizing less complex models with higher-quality inputs. The details regarding the layers are given as follows:

- The input layer consists of neurons that take the 48×48×3 input samples, perform several operations, and transfer the outputs to the subsequent layer.
- While keeping the sample size intact, 3×3 and 5×5 filters convolve the pixels with a single stride and padding of the exact boundary pixels.
- A 2D pooling layer is used to down-sample the input after each convolutional layer. This layer is used to reduce the overfitting and computational overhead of the model by exploring the dominant features that are rotation and position invariant.
- The batch normalization technique is used to normalize the values after each iteration to smooth the convergence procedure.
- A dropout ratio of 20% is used as a regulation technique.
- Two Fully Connected (FC) layers with 128 and 64 nodes perform the classification.
- The nonlinear Rectified Linear Unit (ReLU) function is used for the inner layers to map the net outputs into desired bounded criteria. For the outer layer, the SoftMax activation function is used to improve the final classification of the model based on the highest probability.

For an interested reader, the details regarding the architecture of the current network are also presented in Table 2.

Table 2

Architecture of the current CNN model

Layer Number	Layer Type	Output Dimension	Number of Parameters
1	Convolutional Layer (Conv2D)	48×48×64	4864
2	Convolutional Layer (Conv2D)	48×48×64	102464
3	Batch Normalization	48×48×64	256
4	Max Pooling (3D)	24×24×64	0
5	Convolutional Layer (Conv2D)	24×24×128	204928
6	Convolutional Layer (Conv2D)	24×24×128	409728
7	Convolutional Layer (Conv2D)	24×24×128	409728
8	Batch Normalization	24×24×128	512
9	Max Pooling (3D)	12×12×128	0
10	Convolutional Layer (Conv2D)	12×12×256	295168
11	Convolutional Layer (Conv2D)	12×12×256	590080
12	Convolutional Layer (Conv2D)	12×12×256	590080
13	Batch Normalization	12×12×256	1024
14	Max Pooling (3D)	6×6×256	0
15	Flatten	9216	0
16	Dense	128	1179776
17	Batch Normalization	128	512
18	Activation	128	0
19	Dropout	128	0
20	Flatten	128	0
21	Dense	64	8256
22	Batch Normalization	64	256
23	Activation	64	0
24	Dropout	64	0
25	Dense	7	455
26	Activation	7	0
		Total Number of Parameters: 3,798,087	
		Total Trainable Parameters: 3,796,807	
Total Number of Parameters: 1,280			

3.5. *Transfer Learning*

Transfer learning is mainly referred to as the technique in which a neural network model is trained from a base model with a similar functionality [47]. Although using a pre-trained model leads to fewer training epochs, the model must be exposed to sufficient sample counts to remove any potential mispredictions inherited from the base model.

In this direction, the current study utilized a transfer learning strategy such that the occluded model is trained on top of a pre-trained non-occluded network. There are at least two significant benefits to undergoing such a decision. First, an occluded model trained on top of a non-occluded network is more accurate in classifying human expressions since it inherits a wide range of facial features from the base model, which can improve the predictions odds with fewer samples. Second, utilizing a pre-trained model trained accelerates the training process. The noted issue can be considered an exploitable advantage, especially nowadays, in which many pre-trained models are available. In the following investigations, the effectiveness of the described scenario is assessed.

4. Results

The results are discussed in four sections. Firstly, the assumptions regarding the occlusion and its differences compared with prior studies are described. Secondly, the evaluation metrics used in this study are discussed. Thirdly, the limitation involved in the current study is stated. Lastly, a comparison regarding the accuracy of the present work against related studies is conducted.

4.1. *Occlusion Assumption*

Among images available in the FER-2013 dataset, the facial area corresponding to a limited number of samples is covered due to the hand position, facial orientations, and using sunglasses. In this regard, studies such as [26] consider FER-2013 data as Naturally Occluded Imaging Samples (NOIS) and later evaluate the results through such an assumption.

On the contrary, the current study assumes all the FER-2013 samples are non-occluded, and the occlusion is manually applied through an artificial mask. On average, the occlusion used in this study covers 55% to 70% of the facial area, which can be considered a significantly hard constraint on the solution procedure since the CNN model relies only on the eye and eyebrows regions to classify the samples while locations corresponding to the mouth and jaw that are essential to detect facial expressions such as happiness are entirely unavailable [48]. Such a constraint can highlight the necessity of transfer learning for occluded FER applications.

4.2. *Performance Analysis Metrics*

An accurate performance analysis is more likely to yield highly accurate models. In this regard, the model's accuracy should be strictly monitored during the training and testing session. However, since the inhomogeneity of sample distribution in the FER-2013 dataset is significant, the Accuracy metric alone is not an appropriate metric to decide the maturity of the trained model. As a resolution, the performance of classifiers with inhomogeneous datasets is mainly determined by evaluating the combination of Accuracy, Sensitivity, Precision, and F1-score metrics.

Moreover, it is vital to achieve a single set of performance metrics to evaluate the functionality of the current model against prior works. The noted issue is more sensitive as the number of categories in a multi-class classification problem increases since it is required to compare the performance of a particular classifier against another one in a general format. In the current study, the overall performance metrics are obtained by calculating the weighted average according to the distribution rate of samples in each class.

4.3. *Limitations*

There are several limitations involving in the current study. First, no public dataset is available for researchers who intend to work on occluded FER. Although the presented research has manually applied the occlusion to remove such limitations, a sufficiently large dataset containing occluded samples could further improve the classification performance of the model.

Next, the computational resources used to perform training were limited; therefore, not much hyperparameter tuning was conducted. Utilizing a higher computational resource, the results obtained from this study can be further improved.

4.4. *Comparison with Related Studies*

Although the accuracy value corresponding to studies carried out on FER-2013 is between 55% to 76%, it does not necessarily imply that such studies are not adequately developed. According to [38], the human-level accuracy on the FER-2013 dataset is around 66%, which means that

models with accuracy above 62% in FER-2013 are equivalent to those with an accuracy rate above 95% on datasets, such as CK, CK+, JAFFE, IMFDB, and FEI. As a result, models built using FER-2013 are more prone to noise.

Table 3 compares the accuracy of the current model against some related works built on the FER-2013 dataset in recent years. Based on the results, almost all studies by CNN models outperform the traditional ML models, such as [12] with at least a 10% margin.

The related studies presented in Table 3 can be divided into two main approaches. The first group of studies, such as [13, 28, 30], can be categorized as ones that applied manual filtering, while the latter group, such as [24-26], focused on developing complicated hybrid models to optimize the classification. Based on their results, it can be established that both techniques have successfully improved the FER to some extent, depending on their selected model. However, the results also indicate that the second group dealt with more complications in training their sophisticated models properly. Additionally, as the model complexity grows, more extensive computational resources are needed to realize its full potential. Issues such as sample insufficiency, over-fitting, vanishing gradient, and others are observed in related studies using advanced models such as ResNet and GoogleNet training [31].

Among the noted group, the current study can be categorized among the first group in which manual optimizations such as filtering is adopted. Considering that the modified VGG16 classifier used in this study has achieved an accuracy of 71% without using pre-trained models in the NOIS scenario, it can be concluded that utilizing a multi-channel structure of HOG, Haar, and grayscale is a suitable representation of FER samples and the proposed method is sufficiently accurate for further development.

Lastly, since no occlusion was involved in the related works built on top of the FER-2013 dataset, the non-occluded *Model A* is utilized for the comparison. However, the results corresponding to the occlusion are discussed in the following investigations.

Table 3

Comparison to previous studies

Models	Year	Filters	Classifier	Accuracy
Current study	2023	HOG-Haar	VGG16	71%
[38]	2018	-	None (Human Accuracy)	68%
[12]	2013	HOG	SVM	57%
[13]	2019	Haar	CNN	60%
[15]	2017	-	CNN with Global Average Pooling	66%
[31]	2018	-	GoogleNet	65%
[23]	2020	-	CNN with variable kernel size	65%
[24]	2019	-	Ensemble MLCNN	74%
[25]	2021	-	Attentional CNN	70%
[26]	2023	-	TL-STN	73%
[28]	2017	SIFT	CNN with Dense SIFT	73%
[30]	2020	HOG	CNN	74%

5. Discussion

The discussions are divided into four sections. Firstly, the impact of the HOG and Haar filters is studied individually, and their results are compared against the raw-input case in which no preprocessing is carried out. The noted investigation demonstrates the influence of each filter on feature extraction, thus, reflecting the potential of utilizing application-based multi-channel structure for the input samples instead of raw grayscale or RGB format. Second, the impact of the occlusion on the classification performance of FER models is investigated. The noted analysis reveals the sensitivity of non-occluded models to the locations corresponding to the jaw and mouth. This sensitivity determines the upper-bound performance of the occluded models when trained on top of an occluded model (transfer learning). Also, facial expressions more sensitive to mouth and jaw locations would be detected. Third, the impact of the multi-channel technique on improving the performance of

the FER classifiers is investigated for both occluded and non-occluded samples. Lastly, the effect of transfer learning on the performance of the occluded classifier is investigated. As discussed earlier, many pre-trained models for non-occluded FER are already available to improve the accuracy of occluded models considering the described strategy is found beneficial.

5.1. Impact of Processed Samples on FER

The complexity of famous CNN architectures is due to the excessive number of feature extraction layers required to highlight influential features. As an adverse side-effect, issues such as vanishing gradients, sensitivity to initial weights, over-fitting, and high computational load are also escalating as the model intricacy and the solution procedure.

The current study suggests overcoming the described obstacle using processed data as the initial input of a relatively less complicated neural network model without requiring a significant computational load. In this regard, several filters are selected to process the FER-2013 samples.

As the initial experiment, the performance of the HOG filter is investigated on the classifier's performance, and the results are presented as *Model A* outputs. Based on the outputs shown in Table 4, using the HOG filter decreases all the performance metrics corresponding to the classifier by almost 3~6% since it essentially removes a wide range of features at the cost of highlighting a few. Therefore, the predictions corresponding to the classifier decreased from 67% to 62% after applying the HOG filter. However, the noted observance should not imply that the HOG filter is against the direction of the described arguments since HOG is just one of the layers selected for the multi-channel structure.

Next, the described procedure is carried out for the Haar-wavelet filter. The results indicate a minor improvement (1~2%) in the classification performance, which means that the ratio of impactful features highlighted by the Haar-wavelet filter is slightly more than the ones removed by it. Similar to the HOG scenario, such slight improvement must not be interpreted such that Haar-wavelet does not produce data loss on the original sample.

The simultaneous impact of multi-filtering (multi-channel) is also investigated under the *Model C* label. Since the multi-channel structure contains the original grayscale image that poses all the significant features of specific data; therefore, selecting a low-performance filter would reduce the accuracy lower than what is already achieved without using any filter (*Base Model)*. The reason is that the original grayscale image functions similarly to the skip-state in recurrent networks, which maintains all the lost features independent of the filter operations. As a result, the multi-channel structure brings the benefits of both HOG and Haar-wavelet filters without restricting the model from using either. Consequently, the multi-channel design or the *Model C* scenario produces the highest accuracy score among the tested scenarios.

Table 4

Comparing the FER performance based on the accuracy of suggested filters

Scenario	Channel Type	Occlusion	Filter Type	Accuracy	Sensitivity	Precision	F1-score
Base Model	Single-Channel	No	None	67%	67%	70%	67%
Model A	Single-Channel	No	HOG	62%	62%	63%	61%
Model B	Single-Channel	No	Haar-wavelet	69%	69%	68%	57%
Model C	Multi-Channel	No	HOG, Haar, Grayscale	71%	71%	73%	71%

5.2. Influence of Occlusion on FER

As discussed earlier, the sample occlusion carried out in this study covers 55% of the facial areas at a minimum. Therefore, a valuable portion of information loss related to the mouth, lips, and jaw areas is expected even before undertaking the classification process.

This issue is evaluated by training occluded and non-occluded samples without applying filters. Based on the results presented in Table 5, the mispredictions are increased by almost 25%, demonstrating the impact of face features corresponding to the covered area. Besides the accuracy, other metrics, such as sensitivity, are also reduced from 67% to 43%, indicating the model's ineffectiveness in classifying facial expressions after occlusion is carried out.

Next, the FER classes corresponding to the highest accuracy loss are identified. Based on the results shown in Table 6, it is observed that the disgust, sad, and neutral classes are the ones suffering from the highest portion of feature loss, which can be justified through two reasons.

First, since the noted classes have the fewest available samples, the CNN model might not be not sufficiently developed to produce accurate predictions based on a limited number of upper-face features when applying the medical mask. As a good example, the disgust class has only a 1% portion of the overall data, which reduces the model's robustness as soon as the occlusion removes some portion features. By doing so, the sensitivity of the disgust class is reduced from 63% to 11% when the medical mask is applied.

Second, facial expressions dependent on lower face features are more prone to misprediction after applying the mask. For instance, facial expressions such as sadness and neutrality cannot be effectively detected when the mouth and jaw areas are all covered. Therefore, the model can easily mispredict the noted expressions and produce inaccurate results.

In conclusion, the described issues are responsible for decreasing the accuracy of the occluded samples. In this regard, the current study investigates the impact of utilizing multi-channel inputs followed by a transfer learning strategy to reflect more features on the model.

Table 5

Influence of occlusion on accuracy metrics

Scenario	Channel Type	Occlusion	Filter Type	Accuracy	Sensitivity	Precision	F1-score
Base Model	Single-Channel	No	None	67%	67%	70%	67%
Base Model	Single-Channel	Yes	None	43%	43%	48%	39%

Table 6

Influence of occlusion on accuracy of facial expression classes

Facial Expression	Non-Occluded Base Model			Occluded Base Model		
	Sensitivity	Precision	F1-score	Sensitivity	Precision	F1-score
Angry	75%	53%	63%	63%	28%	39%
Disgust	63%	63%	63%	11%	66%	20%
Fear	39%	63%	48%	30%	26%	28%
Happy	89%	85%	87%	67%	57%	61%
Sad	70%	47%	56%	15%	46%	21%
Surprise	77%	83%	80%	72%	53%	61%
Neutral	45%	81%	58%	11%	62%	18%

5.3. Impact of Multi-Channel Structure on FER

Although it is already established that multi-channel structure can effectively improve the predictions of non-occluded samples, the current section investigates the extent of the noted approach for occluded data in more detail. In this regard, the multi-channel model is trained based on occluded samples, and the results are compared against the occluded *Base Model* presented in Table 7. Based on the outputs, all accuracy metrics are increased by almost 10%, which can be considered a perceptible improvement over the *Base Model*. For example, the sensitivity value of the *Base Model* is increased from 43% to 52.8% after the multi-channel structure is applied instead of the conventional grayscale.

More information regarding the accuracy metrics of each facial expression class is presented in Table 8. As expected, all accuracy scores increased between 5% to 15%, demonstrating that the current filters have successfully highlighted significant features for the classifier.

While the described technique optimizes the feature extraction involved in facial recognition applications, they are not expected to fully compensate for the portion of features lost due to the occlusion process. For instance, even though multi-channel inputs increased the F1-score by 5%,

the corresponding value of 25% is comparably lower than the other classes, such as fear, in which the F1-score is increased from 28% to 62% after the multi-channel method is employed. Similar problems can be observed for facial expressions such as disgust, in which lower face regions are vital.

Nonetheless, it can be concluded that combining transfer learning and the multi-channel filters demonstrates a promising increase in the prediction accuracy of occluded FER classifiers since the model is exposed to a broader range of features which is beneficial to classify samples where a portion of facial areas are concealed from the model.

Table 7

Influence of multi-channel structure on accuracy metrics

Scenario	Channel Type	Filter Type	Accuracy	Sensitivity	Precision	F1-score
Base Model	Single-Channel	None	43%	43%	48%	39%
Model D	Multi-Channel	HOG, Haar, Grayscale	53%	53%	54%	50%

Table 8

Influence of multi-channel structure on occluded FER (Model D)

Facial Expression	Sensitivity	Precision	F1-score
Angry	72%	35%	47%
Disgust	16%	66%	25%
Fear	62%	61%	62%
Happy	76%	60%	67%
Sad	25%	42%	31%
Surprise	71%	64%	67%
Neutral	41%	54%	47%

5.4. FER Transfer Learning

For the final experiment, a pre-trained model developed based on non-occluded samples is considered the initial case to train the occluded model. To do so, *Model D*, which has the highest accuracy score, is selected from the previous investigation, and the occluded samples are used to train the network. As presented in Table 9, since the occlusion imposed by medical masks is applied on the lower facial parts, standard models in which searching for facial features is accomplished are most likely to fail in detecting the correct facial expressions. Nonetheless, by employing transfer learning, the detection accuracy is improved up to 23% compared with the occluded *Base Model*, even though the facial mask covered 55% to 65% of the facial areas. Also, transfer learning improves accuracy by up to 13% compared with the multi-channel scenario (*Model E*). As a result, it is safe to elaborate that utilizing pre-trained models based on non-occluded samples reveals a broader range of features and their relation to one another, which to some certain extent, compensates for the ones removed during the occlusion process.

Compared with models such as [15, 23, 31] that achieved an accuracy rate between 60% to 65% on the non-occluded samples, the current study achieves an accuracy of 66% even when the mask removes half of the facial features on occluded data which demonstrates the impact of training occluded models on top of non-occluded samples.

The results are further investigated by evaluating how much the classification of each class improved after using the transfer learning technique. Based on the results, it is observed that the disgust and sad classes have the highest increase. For instance, the sensitivity of the sad class has improved by more than 40% compared with the occluded *Base Model*. Even compared with the multi-channel structure, all the accuracy metrics were increased by at least 25% for the sad class and between 15% to 40% for both the disgust and the neutral categories. As a result, it can be concluded that utilizing pre-trained non-occluded models (transfer learning) is an effective technique for developing high-resolution models for FER applications.

Table 9

Influence of transfer learning on accuracy metrics

Scenario	Pretrained	Filter Type	Accuracy	Sensitivity	Precision	F1-score
Base Model	No	None	43%	43%	48%	39%
Model E	Yes	HOG, Haar, Grayscale	66%	66%	72%	67%

Table 10

Influence of transfer learning on occluded FER (Model E)

Facial Expression	Sensitivity	Precision	F1-score
Angry	82%	45%	58%
Disgust	81%	81%	81%
Fear	61%	74%	67%
Happy	60%	88%	71%
Sad	70%	53%	60%
Surprise	84%	79%	81%

6. Future Work

There are three separate research paths following the current study. First, the present study selects HOG, Haar wavelet, and grayscale as the multi-channel structure to highlight the facial features; however, there are no limitations on the filter types, and the impact of other suitable filters can be investigated on the feature extraction performance can be evaluated. Similarly, the number of multi-channel layers can be considered an effective hyperparameter that can be optimized during the training phase since a redundant increase in multi-channel layers could potentially slow down the training process.

Second, the current study used VGG16 architecture with a few modifications to perform the FER process. While using more advanced models more likely leads to a more accurate classification, issues such as computational load escalate much faster for such models as the number of multi-channel layers increases. Fortunately, the multi-channel structure supports parallelization on GPU; therefore, it is a suitable method for future development.

Lastly, increasing the number of training samples, especially for facial expressions with few available images, such as disgust and surprise, can improve the model predictions. Hopefully, occluded FER databases will be available in the near future.

7. Conclusion

The current study investigated a novel multi-channel structure of images to improve the accuracy of facial expressions from occluded samples. To do so, the Histogram of Oriented Gradients (HOG) and Haar wavelet, well-established filters for image-processing tasks, were selected alongside the original grayscale sample to highlight the significant facial features and construct the multi-channel kernel. Using the VGG16 classifier on the FER-2013 dataset, the proposed method has obtained an accuracy rate of 71% which can be considered among the high-score models.

By further applying the medical mask using Dlib, the overall accuracy of the base model was decreased by almost 28% since the occlusion covered 45% to 65% of the facial regions containing mouth, lips, and jaw areas. By applying the multi-channel filters for the occluded samples, the accuracy rate has increased by 10%, demonstrating that the noted technique is as effective in highlighting the significant facial features for occluded samples as it was for the non-occluded data.

Lastly, the amount of data lost during occlusion was compensated using pre-trained models based on non-occluded samples. By doing so, a more comprehensive range of facial features and their relation to one another was revealed to the occluded model, which eventually improved the classification performance up to the accuracy rate of 66%. Concerning that almost half of the face was covered by the mask, it can be concluded that utilizing pre-trained non-occluded models (transfer learning) alongside the suggested multi-channel image structure is an effective technique for developing high-resolution models for FER applications.

Authors' Contributions

Mohammad Hossein Zolfagharnasab: Contributed to the conceptualization, methodology design, implementation, data analysis, and interpretation. Assisted in manuscript preparation and editing.

Mohammad Bahrani: Supervised the research project and provided critical revisions to the manuscript. Contributed to the design and development of the study framework. Final approval of the version to be published.

Masood Hamed Saghayan: Involved in methodology, implementation, analysis, and visualization. Assisted in manuscript editing, and revision.

Fatemeh Sadat Masoumi: Conducted the literature review, prepared initial drafts, and managed formatting and references.

Declaration

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence this paper. The contributors are Mohammad Hossein Zolfagharnasab, Mohammad Bahrani, Masood Hamed Saghayan, and Fatemeh Sadat Masoumi.

Transparency Statement

The datasets generated during and/or analyzed during the current study are available in the Harvard Dataverse repository, *https://doi.org/10.7910/DVN/IQJTOT*

Acknowledgments

The authors gratefully acknowledge the insightful comments and constructive recommendations provided by the unknown referee during the review process, which have enhanced the presentation quality of our work.

Declaration of Interest

The authors declare that they have no conflict of interest. The authors also declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Abbreviations

ANN	Artificial Neural Network
CNN	Convolutional Neural Network
DL	Deep Learning
DWT	Discrete Wavelet Transform
FC	Fully Connected
FER	Facial Expression Recognition
FER-CNN	Facial Expression Recognition with CNN
FNR	False Negative Rate
FPR	False Positive Rate
RNN	Recurrent Neural Network
PCA	Principle component analysis
ReLU	Rectified Linear Unit
SVM	Support vector Machine
TNR	True Negative Rate
GF	Gabor Filters
LBP	Local Binary Pattern Operator
LGBP	Gabor Binary Pattern
kNN	k-Nearest Neighbor
TPR	True Positive Rate
SBNN	Searched Based Neural Networks
NOIS	Naturally Occluded Imaging Samples
ICML	International Conference on Machine Learning
ACNN	Attentional Convolutional Neural Networks
SIFT	Scale Invariant Feature Transform
HOG	Histogram of Gradients
DWT	Discrete Wavelet transform
KNN	K-nearest neighbor
LSTM	Long short-time memory
HOG	Histogram of Gradients
DWT	Discrete Wavelet transform

Funding

The current research is conducted by no others rather than the affiliated researchers and is supported by no organization or funding.

Ethical Considerations

This research adheres to ethical guidelines by ensuring the FER-2013 dataset used is publicly accessible and free from privacy violations. Artificial occlusion for this study was applied computationally, ensuring no misuse of sensitive data.

References

[1] H. M. Grillo and M. Enesi, "Impact, importance, types, and use of non-verbal communication in social relations,"

Linguistics and Culture Review, vol. 6, 2022, doi: 10.21744/lingcure.v6ns3.2161.

[2] L. E. Ishii, J. C. Nellis, K. D. Boahene, P. Byrne, and M. Ishii, "The Importance and Psychology of Facial Expression," 2018, doi: 10.1016/j.otc.2018.07.001.

[3] C. Gan, J. Xiao, Z. Wang, Z. Zhang, and Q. Zhu, "Facial expression recognition using densely connected convolutional neural network and hierarchical spatial attention," *Image Vis Comput*, vol. 117, 2022, doi: 10.1016/j.imavis.2021.104342.

[4] Vandana and N. Marriwala, "Facial Expression Recognition Using Convolutional Neural Network," in *Lecture Notes in Networks and Systems*, 2022, doi: 10.1007/978-981-16-7018-3_45.

[5] A. Bremhorst, D. S. Mills, H. Würbel, and S. Riemer, "Evaluating the accuracy of facial expressions as emotion indicators across contexts in dogs," *Anim Cogn*, vol. 25, no. 1, 2022, doi: 10.1007/s10071-021-01532-1.

[6] H. A. H. Mahmoud, N. S. Alghamdi, and A. H. Alharbi, "Real time feature extraction deep-cnn for mask detection," *Intelligent Automation and Soft Computing*, vol. 31, no. 3, 2022, doi: 10.32604/IASC.2022.020586.

[7] F. Pazhoohi, L. Forby, and A. Kingstone, "Facial masks affect emotion recognition in the general population and individuals with autistic traits," *PLoS One*, vol. 16, no. 9, 2021, doi: 10.1371/journal.pone.0257740.

[8] M. Kuzu Kumcu, S. Tezcan Aydemir, B. Ölmez, N. Durmaz Çelik, and C. Yücesan, "Masked face recognition in patients with relapsing-remitting multiple sclerosis during the ongoing COVID-19 pandemic," *Neurological Sciences*, vol. 43, no. 3, 2022, doi: 10.1007/s10072-021-05797-9.

[9] F. Yan, N. Wu, A. M. Iliyasu, K. Kawamoto, and K. Hirota, "Framework for identifying and visualising emotional atmosphere in online learning environments in the COVID-19 Era," *Applied Intelligence*, vol. 52, no. 8, 2022, doi: 10.1007/s10489-021-02916-z.

[10] R. Gonzales-Martinez, J. Machacuay, P. Rotta, and C. Chinguel, "Hyperparameters Tuning of Faster R-CNN Deep Learning Transfer for Persistent Object Detection in Radar Images," *IEEE Latin America Transactions*, vol. 20, no. 4, 2022, doi: 10.1109/TLA.2022.9675474.

[11] S. Kim, K. Lee, M. Lee, J. Lee, T. Ahn, and J. T. Lim, "Evaluation of saturation changes during gas hydrate dissociation core experiment using deep learning with data augmentation," *J Pet Sci Eng*, vol. 209, 2022, doi: 10.1016/j.petrol.2021.109820.

[12] I. J. Goodfellow, "Challenges in representation learning: A report on three machine learning contests," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2013, doi: 10.1007/978-3-642-42051-1_16.

[13] I. Talegaonkar, K. Joshi, S. Valunj, R. Kohok, and A. Kulkarni, "Real Time Facial Expression Recognition using Deep Learning," *SSRN Electronic Journal*, 2019, doi: 10.2139/ssrn.3421486.

[14] O. S. Ekundayo and S. Viriri, "Facial Expression Recognition: A Review of Trends and Techniques," 2021, doi: 10.1109/ACCESS.2021.3113464.

[15] O. Arriaga, M. Valdenegro-Toro, and P. G. Plöger, "Real-time convolutional neural networks for emotion and gender classification," in *ESANN 2019 - Proceedings, 27th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, 2019. [Online]. Available: https://www.esann.org/sites/default/files/proceedings/legacy/es2019-157.pdf. [Online]. Available: https://www.esann.org/sites/default/files/proceedings/legacy/es2019-157.pdf

[16] H. Ge, Z. Zhu, Y. Dai, B. Wang, and X. Wu, "Facial expression recognition based on deep learning," *Comput Methods Programs Biomed*, vol. 215, 2022, doi: 10.1016/j.cmpb.2022.106621.

[17] T. Gwyn, K. Roy, and M. Atay, "Face recognition using popular deep net architectures: A brief comparative study," *Future Internet*, vol. 13, no. 7, 2021, doi: 10.3390/fi13070164.

[18] A. T. Kabakus, "PyFER: A Facial Expression Recognizer Based on Convolutional Neural Networks," *IEEE Access*, vol. 8, 2020, doi: 10.1109/ACCESS.2020.3012703.

[19] S. Li and W. Deng, "Deep Facial Expression Recognition: A Survey," *IEEE Trans Affect Comput*, vol. 13, no. 3, 2022, doi: 10.1109/TAFFC.2020.2981446.

[20] H. Dino, "Facial Expression Recognition based on Hybrid Feature Extraction Techniques with Different Classifiers," *TEST Engineering & Management*, vol. 83, no. 22319, 2020. [Online]. Available: https://www.researchgate.net/publication/342317981_Facial_Expression_Recognition_based_on_Hybrid_Feature_Extraction_Techniques_with_Different_Classifiers.

[21] S. B. Sukhavasi, S. B. Sukhavasi, K. Elleithy, A. El-Sayed, and A. Elleithy, "Deep Neural Network Approach for Pose, Illumination, and Occlusion Invariant Driver Emotion Detection," *Int J Environ Res Public Health*, vol. 19, no. 4, 2022, doi: 10.3390/ijerph19042352.

[22] H. Elaggoune, M. Belahcene, and S. Bourennane, "Hybrid descriptor and optimized CNN with transfer learning for face recognition," *Multimed Tools Appl*, vol. 81, no. 7, 2022, doi: 10.1007/s11042-021-11849-1.

[23] A. Agrawal and N. Mittal, "Using CNN for facial expression recognition: a study of the effects of kernel size and number of filters on accuracy," *Visual Computer*, vol. 36, no. 2, 2020, doi: 10.1007/s00371-019-01630-9.

[24] H. D. Nguyen, S. H. Kim, G. S. Lee, H. J. Yang, I. S. Na, and S. H. Kim, "Facial Expression Recognition Using a Temporal Ensemble of Multi-Level Convolutional Neural Networks," *IEEE Trans Affect Comput*, vol. 13, no. 1, 2022, doi: 10.1109/TAFFC.2019.2946540.

[25] S. Minaee, M. Minaei, and A. Abdolrashidi, "Deep-emotion: Facial expression recognition using attentional convolutional network," *Sensors*, vol. 21, no. 9, 2021, doi: 10.3390/s21093046.

[26] J. Kim and D. Lee, "Facial Expression Recognition Robust to Occlusion and to Intra-Similarity Problem Using Relevant Subsampling," *Sensors*, vol. 23, no. 5, 2023, doi: 10.3390/s23052619.

[27] M. N. Ab Wahab, A. Nazir, A. T. Z. Ren, M. H. M. Noor, M. F. Akbar, and A. S. A. Mohamed, "Efficientnet-Lite and Hybrid CNN-KNN Implementation for Facial Expression Recognition on Raspberry Pi," *IEEE Access*, vol. 9, 2021, doi: 10.1109/ACCESS.2021.3113337.

[28] T. Connie, M. Al-Shabi, W. P. Cheah, and M. Goh, "Facial expression recognition using a hybrid CNN-SIFT aggregator," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2017, doi: 10.1007/978-3-319-69456-6_12.

[29] M. I. Georgescu, R. T. Ionescu, and M. Popescu, "Local learning with deep and handcrafted features for facial expression recognition," *IEEE Access*, vol. 7, 2019, doi: 10.1109/ACCESS.2019.2917266.

[30] S. Jaiswal and G. C. Nandi, "Robust real-time emotion detection system using CNN architecture," *Neural Comput*

Appl, vol. 32, no. 15, 2020, doi: 10.1007/s00521-019-04564-4.

[31] P. Giannopoulos, I. Perikos, and I. Hatzilygeroudis, "Deep learning approaches for facial emotion recognition: A case study on FER-2013," in *Smart Innovation, Systems and Technologies*, 2018, doi: 10.1007/978-3-319-66790-4_1.

[32] C. Pramerdorfer and M. Kampel, "Facial Expression Recognition using Convolutional Neural Networks: State of the Art," *ArXiv*, 2016. [Online]. Available: https://arxiv.org/abs/1612.02903.

[33] X. Wang, J. Huang, J. Zhu, M. Yang, and F. Yang, "Facial expression recognition with deep learning," in *ACM International Conference Proceeding Series*, 2018, doi: 10.1145/3240876.3240908.

[34] S. I. Saleem and A. M. Abdulazeez, "Hybrid trainable system for writer identification of arabic handwriting," *Computers, Materials and Continua*, vol. 68, no. 3, 2021, doi: 10.32604/cmc.2021.016342.

[35] J. Sahar Zafar, A. Fayyaz, G. Subhash, I. A. Kandhro, A. Khan, and A. Zaidi, "Facial Expression Recognition with Histogram of Oriented Gradients using CNN," *Indian J Sci Technol*, vol. 12, no. 24, 2019, doi: 10.17485/ijst/2019/v12i24/145093.

[36] R. Azmi and S. Yegane, "Facial expression recognition in the presence of occlusion using local Gabor binary patterns," in *ICEE 2012 - 20th Iranian Conference on Electrical Engineering*, 2012, doi: 10.1109/IranianCEE.2012.6292452.

[37] H. Zolfagharnasab *et al.*, "A regression model for predicting shape deformation after breast conserving surgery," *Sensors (Switzerland)*, vol. 18, no. 1, 2018, doi: 10.3390/s18010167.

[38] S. Saeed *et al.*, "Empirical evaluation of SVM for facial expression recognition," *International Journal of Advanced Computer Science and Applications*, vol. 9, no. 11, 2018, doi: 10.14569/ijacsa.2018.091195.

[39] S. Suwarno and K. Kevin, "Analysis of Face Recognition Algorithm: Dlib and OpenCV," *JOURNAL OF INFORMATICS AND TELECOMMUNICATION ENGINEERING*, vol. 4, no. 1, 2020, doi: 10.31289/jite.v4i1.3865.

[40] T. Mahmud, "Face Detection and Recognition System," in *Lecture Notes in Networks and Systems*, 2021, doi: 10.1007/978-981-16-3153-5_18.

[41] O. Elharrouss, N. Almaadeed, S. Al-Maadeed, A. Bouridane, and A. Beghdadi, "A combined multiple action recognition and summarization for surveillance video sequences," *Applied Intelligence*, vol. 51, no. 2, 2021, doi: 10.1007/s10489-020-01823-z.

[42] T. Zhang, X. Zhang, X. Ke, C. Liu, X. Xu, and X. Zhan, "HOG-ShipCLSNet: A Novel Deep Learning Network with HOG Feature Fusion for SAR Ship Classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, 2022, doi: 10.1109/TGRS.2021.3082759.

[43] S. Ganesan, R. Dr, and S. J. Dr, "Prediction of Autism Spectrum Disorder by Facial Recognition Using Machine Learning," *Webology*, vol. 18, no. 02, 2021, doi: 10.14704/web/v18si02/web18291.

[44] A. B. Nassif, A. M. Darya, and A. Elnagar, "Empirical Evaluation of Shallow and Deep Learning Classifiers for Arabic Sentiment Analysis," *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 21, no. 1, 2022, doi: 10.1145/3466171.

[45] A. Chalehchaleh and A. Khadem, "Diagnosis of Bipolar i Disorder using 1 D-CNN and Resting-State fMRI Data," in *Proceedings of the 5th International Conference on Pattern Recognition and Image Analysis, IPRIA 2021*, 2021, doi: 10.1109/IPRIA53572.2021.9483574.

[46] H. M. Rizwan Iqbal and A. Hakim, "Classification and Grading of Harvested Mangoes Using Convolutional Neural Network," *International Journal of Fruit Science*, vol. 22, no. 1, 2022, doi: 10.1080/15538362.2021.2023069.

[47] M. H. Zolfagharnasab and S. Damari, "A Comparative Analysis of Machine Learning Models in News Categorization," *U. Porto Journal Of Engineering*, 2024, doi: 10.24840/2183-6493_0010-003_002464.

[48] D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation," *BMC Genomics*, vol. 21, no. 1, 2020, doi: 10.1186/s12864-019-6413-7.