Facial Emotion Recognition Using Conventional Machine Learning and Deep Learning Methods: Current Achievements, Analysis and Remaining Challenges

Amjad Rehman Khan

Artificial Intelligence & Data Analytics (AIDA) Lab, CCIS Prince Sultan University, Riyadh 11586, Saudi Arabia; arkhan@psu.edu.sa

Abstract: Facial emotion recognition (FER) is an emerging and significant research area in the pattern recognition domain. In daily life, the role of non-verbal communication is significant, and in overall communication, its involvement is around 55% to 93%. Facial emotion analysis is efficiently used in surveillance videos, expression analysis, gesture recognition, smart homes, computer games, depression treatment, patient monitoring, anxiety, detecting lies, psychoanalysis, paralinguistic communication, detecting operator fatigue and robotics. In this paper, we present a detailed review on FER. The literature is collected from different reputable research published during the current decade. This review is based on conventional machine learning (ML) and various deep learning (DL) approaches. Further, different FER datasets for evaluation metrics that are publicly available are discussed and compared with benchmark results. This paper provides a holistic review of FER using traditional ML and DL methods to highlight the future gap in this domain for new researchers. Finally, this review work is a guidebook and very helpful for young researchers in the FER area, providing a general understating and basic knowledge of the current state-of-the-art methods, and to experienced researchers looking for productive directions for future work.

Keywords: facial expressions; facial emotion recognition (FER); technological development; healthcare; security; deep learning and traditional classification methods

1. Introduction

Facial emotions and their analysis play a vital role in non-verbal communication. It makes oral communication more efficient and conducive to understanding the concepts [1,2]. It is also conducive to detecting human attention, such as behavior, mental state, personality, crime tendency, lies, etc. Regardless of gender, nationality, culture and race, most people can recognize facial emotions easily. However, a challenging task is the automation of facial emotion detection and classification. The research community uses a few basic feelings, such as fear, aggression, upset and pleasure. However, differentiating between many feelings is very challenging for machines [3,4]. In addition, machines have to be trained well enough to understand the surrounding environment—specifically, an individual's intentions. When machines are mentioned, this term includes robots and computers. A difference is that robots involve communication abilities to a more innovative extent since their design consists of some degree of autonomy [5,6]. The main problem is classifying people's emotions is variations in gender, age, race, ethnicity and image quality or videos. It is necessary to provide a system capable of recognizing facial emotions with similar knowledge as possessed by humans. Recently, FER has become an emerging field of research, particularly for the last few decades. Computer vision techniques, AI, image processing and ML are widely used to expand effective automated facial recognition systems for security and healthcare applications [7-10].

Face detection is the first step of locating or detecting face(s) in a video or single image in the FER process. The images do not consist of faces only, but instead present
with complex backgrounds. Indeed, human beings can easily predict facial emotions and other facial features of an image, but these are difficult tasks for machines without excellent training [11,12]. The primary purpose of face detection is to separate face images from the background (non-faces). Some face detection domains are gesture recognition, video surveillance systems, automated cameras, gender recognition, facial feature recognition, face recognition, tagging and teleconferencing [13,14]. These systems need first to detect faces as inputs. There is a color sensor for image acquisition that captures color images everywhere. Hence, current face recognition techniques depend heavily on grayscale, and there are only a few techniques that are able to operate with color images. To achieve better performance, these systems either implement window-based or pixel-based techniques, which are the key categories of strategies for facial recognition. The pixel-based method lags in separating the face from the hands of another skin region of the person [15,16].

In contrast, the window-based method loses the capacity to view faces from various perspectives. Model matching approaches are the most frequently used techniques for FER, including face detection. In contrast, the window-based approach cannot view faces from different angles. Currently, several state-of-the-art classifiers, such as ANN, CNN, SVM, KNN, and random forest (RF), are employed for different features’ extraction and in the recognition of tumors in healthcare, in biometrics, in handwriting studies and in detecting faces for security measures [17–19].

Main Contributions

Over the past three decades, there has been a lot of research reported in the literature on facial emotion recognition (FER). Nevertheless, despite the long history of FER-related works, there are no systematic comparisons between traditional ML and DL approaches. Kołakowska [20] presented a review whose main focus was on conventional ML approaches. Ghayoumi [21] proposed a short review of DL in FER, and Ko et al. [22] presented a state-of-the-art review of facial emotion recognition techniques using visual data. However, they only provided the differences between conventional ML techniques and DL techniques. This paper presents a detailed investigation and comparisons of traditional ML and deep learning (DL) techniques with the following major contributions:

- The main focus is to provide a general understanding of the recent research and help newcomers to understand the essential modules and trends in the FER field.
- We present the use of several standard datasets consisting of video sequences and images with different characteristics and purposes.
- We compare DL and conventional ML approaches for FER in terms of resource utilization and accuracy. The DL-based approaches provide a high degree of accuracy but consume more time in training and require substantial processing capabilities, i.e., CPU and GPU. Thus, recently, several FER approaches have been used in an embedded system, e.g., Raspberry Pi, Jetson Nano, smartphones, etc.

This paper provides a holistic review of facial emotion recognition using the traditional ML and DL methods to highlight the future gap in this domain for new researchers. Further, Section 1 presents the related background on facial emotion recognition; Sections 2 and 3 contain a brief review of traditional ML and DL. In Section 4, a detailed overview is presented of the FER datasets. Section 5 considers the performance of the current research work, and, finally, the research is concluded in Section 6.

2. Facial Emotion Recognition Using Traditional Machine Learning Approaches

Facial emotions are beneficial for investigating human behavior [23,24] as exhibited in Figure 1. Psychologically, it is proven that the facial emotion recognition process measures the eyes, nose, mouth and their locations.

The earliest approach used for facial emotion intensity estimation was based on distance urged. This approach uses high-dimensional rate transformation and regional volumetric distinction maps to categorize and quantify facial expressions. In videos, most systems use Principal Component Analysis (PCA) to represent facial expression
features [25]. PCA has been used to recognize the action unit to express and establish different facial expressions. Other facial expressions are structured and recognized by mistreatment PCA for providing a facial action unit [26].

Figure 1. Facial emotion recognition (FER) process.

Siddiqi et al. [27] detected and extracted the face portion via the active contour model. The researchers used Chan–Vese and Bhattacharyya’s energy functions to optimize the distance between face and context, and reduce the differences within the face. In addition, noise is reduced using wavelet decomposition, and the geometric appearance features of facial emotions and facial movement features using optical flow are extracted.

There is no need for high computational power and memory for conventional ML methods such as DL methods. Therefore, these methods need further consideration to implement embedded devices that perform classification in real time with low computational power and provide satisfactory results. Accordingly, Table 1 presents a brief summary

Table 1. Summary of the representative conventional FER approaches.

| References          | Datasets                      | Decision Methods               | Features                                          | Emotions Analyzed |
|---------------------|-------------------------------|--------------------------------|--------------------------------------------------|-------------------|
| Varma et al. [28]   | FECR                          | support vector machine (SVM) classifier and HMM | PCA and LDA                                      | Six emotions      |
| Reddy et al. [29]   | CK+                           | support vector machine (SVM) classifier | Haar Wavelet Transform (HWT), Gabor wavelets and nonlinear principal component analysis (NLPCA) | Six emotions      |
| Sajjad et al. [30]  | MMI and JAFFE and CK+         | support vector machine (SVM) classifier | ORB, SIFT, SURF                                  | Seven emotions    |
| Nazir et al. [31]   | CK+ and MMI                   | KNN, SMO and MLP for classification | HOG, DCT                                         | Seven emotions    |
| Zeng et al. [32]    | CK+                           | DSAE                           | LBP, SIFT, Gabor Function and HOG                | Seven emotions    |
| Uddin et al. [33]   | Real-time dataset from visual depth camera | Deep Belief Network (DBN) | MLDP-GDA features                               | Six emotions      |
| Al-Agha et al. [34] | BOSPHORUS                     | Euclidean distance             | Geometric descriptor                             | Four emotions     |
| Ghimire et al. [35] | CK+ and MMI and MUG           | EBGM, KLT and AdaBoost-ELM     | Salient geometric features                       | Seven emotions    |
| Wang and Yang [36]  | BVTKFER and CurtinFaces       | random forest classifier       | LBP                                              | Six emotions      |
### Table 1. Cont.

| References      | Datasets                  | Decision Methods                  | Features                                      | Emotions Analyzed          |
|-----------------|---------------------------|-----------------------------------|-----------------------------------------------|----------------------------|
| Wu et al. [37]  | BBC, SPOS, MMI, UvANEMO   | support vector machine (SVM)      | RSTD, four conventional features—raw pixels, Gabor, HOG and LBP | Two emotions (Smile)       |
| Acevedo et al. [38] | CK+                       | Conditional Random Field (CRF) and KNN | Geometric descriptor                         | Seven emotions             |
| Kim [39]        | CK, JAFFE                 | embedded hidden Markov model (EHMM) | ASM and 2D DCT                                | Seven emotions             |
| Cornejo et al. [25] | CK, JAFFE, MUG          | PCA, LDA, K-NN and SVM            | Gabor wavelets and geometric features         | Seven emotions             |
| Siddiqi et al. [27] | CK, JAFFE, USTCNVIE, Yale, FEI | hidden Markov model (HMM)         | Chan–Vese energy function, Bhattacharyya distance function wavelet decomposition and SWLDA | Six emotions               |
| Chang et al. [40] | CK+                       | Support Vector Regression (SVR)   | feature descriptors, AAMs, Gabor wavelets     | Seven emotions             |

### 3. Facial Emotion Recognition Using Deep-Learning-Based Approaches

Deep learning (DL) algorithms have revolutionized the computer vision field in the current decade with RNN and CNN [41–43]. These DL-based methods are used for feature extraction, recognition and classification tasks. The key advantage of a DL approach (CNN) is to overcome the dependency on physics-based models and reduce the effort required in preprocessing and feature extraction phases [44,45]. In addition, DL methods enable end-to-end learning from input images directly. For these purposes, in several regions, including FER, scene awareness, face recognition and entity recognition, DL-based methods have obtained encouraging results from the state-of-the-art [46,47]. There are generally three layers in a DL-CNN, (1) convolution layer, (2) subsampling layer and (3) FC layer, as exhibited in Figure 2. The CNN takes the image or feature maps as the input, and slides these inputs together with a series of filter banks to produce feature maps that reflect the facial image’s spatial structure. Inside a feature map, the weights of convolutional filters are connected, and the feature map layer inputs are locally connected [48–50]. By implementing one of the most popular pooling approaches, i.e., max pooling, min pooling or average pooling, the second type of layer, called subsampling, is responsible for reducing the given feature maps [51,52]. A CNN architecture’s last FC layer calculates the class probability of an entire input image. Most DL-based techniques can be freely adapted with a CNN to detect emotions.

Li et al. [53] proposed a 3D CNN architecture to recognize several emotions from videos. They extracted deep features and used three benchmark datasets for the experimental evaluation, namely CASME II, CASME and SMIC. Li et al. [54] performed additional face cropping and rotation techniques for feature extraction using a convolutional neural network (CNN). Tests were carried out on the CK+ and JAFFE databases to test the proposed procedure.

A convolution neural network (CNN) with a focus function (ACNN) was suggested by Li et al. [55] that could interpret the occlusion regions of the face and concentrate on more discriminatory, non-occluded regions. A CNN is an end-to-end learning system. First, the different depictions of facial regions of interest (ROIs) are merged. Then, each representation is weighted by a proposed gate unit that calculates an adaptive weight according to the area’s significance. Two versions of ACNN have been developed in separate ROIs: patch-based CNN (pACNN) and global-local-based ACNN (gACNN). Lopes et al. [56] classified human faces into several emotion groups. They used three
different architectures for classification: (1) a CNN with 5 C layers, (2) a baseline with one C layer and (3) a deeper CNN with several C layers. Breuer and Kimmel [57] trained a model using various datasets of FER to classify seven basic emotions. Chu et al. [58] presented multi-level mechanisms for detecting facial emotions by combining temporal and spatial features. They used a CNN architecture for spatial feature extraction and LSTMs to model the temporal dependencies. Finally, they fused the output of both LSTMs and CNNs to provide a per-frame prediction of twelve facial emotions. Hasani and Mahoor [59] presented the 3D Inception-ResNet model and LSTM unit, which were fused to extract temporal and spatial features from the input frames of video sequences. (Zhang et al., [60] and Jain et al. [61] suggested a multi-angle optimal pattern-dependent DL (MAOP-DL) system to address the problem of abrupt shifts in lighting and achieved proper alignment of the feature set by utilizing optimal arrangements centered on multi-angles. Their approach first subtracts the backdrop, isolates the subject from the face images and later removes the facial points’ texture patterns and the related main features. The related features are extracted and fed into an LSTM-CNN for facial expression prediction.

Figure 2. Training process of CNN model for facial emotion recognition.

Al-Shabi et al. [62] qualified and collected a minimal sample of data for a model combination of CNN and SIFT features for facial expression research. A hybrid methodology was used to construct an efficient classification model, integrating CNN and SIFT functionality. Jung et al. [63] proposed a system in which two CNN models with different characteristics were used. Firstly, presence features were extracted from images, and secondly, temporal geometry features were extracted from the temporal facial landmark points. These models were fused into a novel integration scheme to increase FER efficiency. Yu and Zhang [64] used a hybrid CNN to execute FER and, in 2015, obtained state-of-the-art outcomes in FER. They used an assembly of CNNs with five convolution layers for each facet word. Their method imposed transformation on the input image in the training process, while their model produced predictions for each subject’s multiple emotions in the testing phase. They used stochastic pooling to deliver optimal efficiency, rather than utilizing peak pooling (Table 2).

The hybrid CNN-RNN and CNN-LSTM techniques have comparable architectures, as discussed in the previous section and exhibited in Figure 3. In short, CNN-RNN’s simple architecture combines an LSTM with a DL software visual feature extractor, such as the CNN model. The hybrid techniques are, thus, equipped to distinguish emotions from image sequences. Figure 3 indicates that each graphic attribute has been translated to the LSTM blocks and describes a variable or fixed-length vector. Finally, performance is given for the prediction in a recurrent sequence learning module and the SoftMax classifier is used in [58].
### Table 2. Deep learning-based approaches for FER.

| References       | Datasets                          | Methods                                      | Emotions Analyzed      |
|------------------|-----------------------------------|----------------------------------------------|------------------------|
| Li et al. [65]   | CK+ JAFFE                         | convolutional neural network (CNN)           | Seven emotions         |
| Jain et al. [61] | CK+ MMI                           | For the facial label prediction, they used LSTM-CNN | Six emotions           |
| Li et al. [53]   | SMIC CASME CASME II               | 3D flow-based CNN model for video-based micro-expression recognition | Five emotions         |
| Xie et al. [66]  | CK+ JAFFE                         | Attention-based SERD and the MPVS-Net        | Six emotions           |
| Li et al. [55]   | RAF-DB AffectNet SFEW CK+ MMI Oulu-CASIA | CNN with ACNN                              | Seven emotions         |
| Lopes et al. [56]| CK+ JAFFE BU-3DFE                | CNN and specific image preprocessing steps   | Seven emotions         |
| Breuer and Kimmel [57]| CK+ FER2013                      | CNN with transfer learning                  | Seven emotions         |
| Chu et al. [58]  | BP4D                              | CNN is used to extract spatial representations and LSTMs used for temporal dependencies | Twelve emotions        |
| Hasani and Mahoor [59]| CK+ DISFA                        | LSTM units extracted both temporal and spatial relations from facial images; as inputs of the network, they used facial landmark points | Twenty-three primary and compound emotions |
| Zhang et al. [60]| CK+ Oulu-CASIA MMI               | PHRNN and MSCNN                              | Seven emotions         |
| Al-Shabi et al. [62]| CK+ FER-2013                    | Dense SIFT and regular SIFT are merged with CNN features | Seven emotions         |
| Jung et al. [63] | CK+ Oulu-CASIA                   | DTAN and DTGN                                | Seven emotions         |
| Yu and Zhang [64]| FER-2013 SFEW                    | Used an assembly of DCNNs with seven convolution layers | Seven emotions         |

Generally, DL-based methods determine classifiers and features by DNN experts, unlike traditional ML methods. DL-based methods extract useful features directly from training data using DCNNs [67,68]. However, the massive training data are a challenge for facial expressions under different conditions to train DNNs. Furthermore, DL-based methods need high computational power and a large amount of memory to train and test the model compared to traditional ML methods. Thus, it is necessary to decrease the computational time during the inferencing of DL methods.
Figure 3. Demonstrated CNN features used by LSTM for sequence learning; SoftMax is used.

4. Facial Emotion Datasets

Experts from different institutions have generated several datasets to evaluate reported methods for facial expression classification [69,70]. Accordingly, a detailed overview of some benchmark datasets is presented.

- The CK+ Cohen Kanade Dataset:

  The Cohen Kanade database [71] is inclusive in that it consists of subject images of all sexes and races and is open to the public. This dataset consists of seven essential emotions that often include neutral emotions. The images’ resolution is $640 \times 490$ or $640 \times 480$ dimensions, with grayscale (8-bit) existing in the dataset. Approximately 81% of subjects are Euro-American, 13% are Afro American, and approximately 6% are from other races. In the CK+ dataset, the ratio of females is nearly 65%. The dataset consists of 593 images captured from 120 different people, and the age of these people varies from 18 to 30 years.

- Bosphorus dataset

  This dataset consists of 2D and 3D faces for emotion recognition, facial action detection, 3D face reconstruction, etc. There are 105 humans with 4666 faces in different poses in the dataset. This dataset is different from other datasets in the following aspects.

  1. A rich collection of facial emotions is included:
     i. Per person, at least 35 facial emotions are recoded;
     ii. FACS scoring;
     iii. Each third person is a professional actresses/actors.

  2. Systematic head poses are included.

  3. A variety of facial occlusions are included (eyeglasses, hands, hair, moustaches and beards).

- SMIC dataset

  SMIC [54] contains macro- and micro-emotions, but the focus is on macro-emotions; this category includes 164 videos taken from 16 subjects. The videos were captured using a powerful 100 frame per second camera.

- BBC dataset

  The BBC database is from the ‘Spot the fake smile’ test on the BBC website (http://www.bbc.co.uk/science/humanbody/mind/surveys/smiles/) accessed on 17 December.
2021), which consists of 10 genuine and 10 posed smile videos collected with a resolution of $314 \times 286$ pixels at 25 frames/s from seven females and thirteen males.

The image representation of these datasets is shown in Figure 4 such that each row represents an individual dataset.

![Figure 4. Cont.](image-url)
Figure 4. Cont.
Figure 4. Twenty different datasets for facial emotion recognition, including CK+, MMI, JAFFE, SFEW, KDEF, FER-2013, MUG, Yale B face, BOSPHORUS, UvA-Nemo, NUST-NVIE, SPOS, FEI, DIFSA, AFEW, Curtain-face, VT-KFER, Oulu-CASIA, BU-3DFE and BP4D.

5. Performance Evaluation of FER

Performance evaluation based on quantitative comparison is an important technique to compare experimental results [72,73]. Benchmark comparisons on publicly available datasets are also presented. Two different mechanisms are used to evaluate the reported system’s accuracy: (1) cross-dataset and (2) subject-independent. Firstly, a subject-independent task separates each dataset into two parts: validation and training datasets. This process is also known as K-fold cross-validation [74,75]. K-fold cross-validation is used to overcome overfitting and provide insight into how the model will generalize to an unknown, independent dataset.

6. Evaluation Metrics/Performance Parameters

The evaluation metrics include overall accuracy, F-measure, recall and precision [76]. Each evaluation metric is discussed separately below.

- **Precision**: This indicates how accurate or precise a model is, and it yields the actual positive values from a set of predictions by any algorithm. Precision is an excellent measure to indicate when the costs of false positives are relatively very high [77–79].

  \[
  \text{Precision} = \frac{\text{True positive}}{\text{True positive} + \text{False positive}} \quad (1)
  \]

- **Recall**: A recall value computes how many actual positive values a model captures while labeling it positive. This metric is used to select the best model when there is a high cost of false negatives [80,81].
Precision = (True positive)/(True positive + False negative) \hspace{1cm} (2)

- F1 Score: The F1 score is acquired when a balance is needed between precision and recall. F1 is a function of precision and recall \cite{82,83}.
  \[
  F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \hspace{1cm} (3)
  \]
- True Negatives (TN): Samples in the results are graded as not adequately belonging to the required class \cite{84–86}.
- False Negatives (FN): Samples are mistakenly used in data identified as not belonging to the required class \cite{87–90}.
- True Positives (TP): Samples in the data are appropriately categorized as belonging to the desired class \cite{91–93}.
- False Positives (FP): Samples in the data are categorized incorrectly as belonging to the target class \cite{77,94}.

**Comparisons on Benchmark Datasets**

It is challenging to compare different ML and deep-learning-based facial recognition strategies due to the different setups, datasets and machines used \cite{95,96}. However, the latest comparison of different approaches is presented in Table 3.

**Table 3.** Conventional ML-based approaches for FER.

| References       | Datasets          | Accuracy (%)                  | Techniques                                                                 |
|------------------|-------------------|-------------------------------|----------------------------------------------------------------------------|
| Varma et al. \[28\] | FECR              | 98.40 (SVM) 87.50 (HMM)       | PCA and LDA for feature selection and SVM and HMM classifier               |
| Reddy et al. \[29\] | CK+               | 98.00                         | Haar Wavelet Transform (HWT), Gabor wavelets and nonlinear principal component analysis (NLPCA) feature extractor and SVM classifier |
| Sajjad et al. \[30\] | MMI               | 99.1 on MMI accuracy (92)     | ORB SIFT SURF using SVM HOG, DCT features and KNN, SMO and MLP for classification |
|                  | JAFFE             | 99.1 on JAFFE accuracy (90) on CK+ |                                                                         |
|                  | CK+               | 99.6 by using only 32 features with KNN |                                                                         |
| Nazir et al. \[31\] | CK+               | 95.79                         | LBP, SIFT, Gabor Function and HOG features and Deep Sparse Auto-Encoder (DSAE) classifier |
|                  | MMI               |                               | MLDP-GDA features fed to Deep Belief Network (DBN) for prediction         |
|                  | CK+               | 97.80                         | Salient geometric features and EBGM, KLT and AdaBoost-ELM class           |
|                  | MMI               | 77.22                         |                                                                           |
|                  | MUG               | 95.50                         |                                                                           |
| Uddin et al. \[33\] | Real-time dataset from visual depth camera | 96.25                     |                                                                           |
| Ghimire et al. \[35\] | CK+               | 97.80                         |                                                                           |
|                  | MMI               | 77.22                         |                                                                           |
|                  | MUG               | 95.50                         |                                                                           |
| Yang et al. \[96\] | BVT-KFER          | 93.33                         | LBP features and a random forest classifier                               |
|                  | B CurtinFaces     | 92.50                         |                                                                           |
|                  | BBC               | 90.00                         | RSTD, four conventional features—raw pixel, Gabor, HOG and LBP features | Support vector machine (SVM) classifiers |
|                  | SPOS              | 81.25                         |                                                                           |
|                  | MMI               | 92.21                         |                                                                           |
| Wu et al. \[37\]  | UVa-NEMO          | 93.95                         |                                                                           |
Table 3. Cont.

| References          | Datasets          | Accuracy (%) | Techniques                                                                 |
|---------------------|-------------------|--------------|----------------------------------------------------------------------------|
| Acevedo et al. [38] | CK+               | 89.3         | Geometric descriptor and Conditional Random Field (CRF) and KNN classifier  |
|                     |                   | 86.9         |                                                                            |
| Al-Agha et al. [34] | BOSPHORUS         | 85.00        | Geometric descriptor and Euclidean distance is used for the prediction      |
| Kim [39]            | CK                | 77.00        | ASM and 2D DCT features and embedded hidden Markov model (EHMM)            |
|                     | JAFFE             | 64.7         |                                                                            |
| Siddiqi et al. [27] | CK                | 99.33        | Features are extracted by using Chan–Vese energy function, Bhattacharyya |
|                     | JAFFE             | 96.50        | distance function, wavelet decomposition and SWLDA and hidden Markov model (HMM) is used for prediction |
|                     | USTC-NVIE         | 99.17        |                                                                            |
|                     | Yale              | 99.33        |                                                                            |
|                     | FEI               | 99.50        |                                                                            |
| Cornejo et al. [25] | CK+               | 99.78        | Geometric + PCA + LDA + KNN                                              |
|                     | JAFFE             | 99.17        | Geometric + PCA + LDA + K – NN                                           |
|                     | MUG               | 99.94        | Geometric + PCA + LDA + K – NN                                           |
| Chang et al. [40]   | CK+               | Mean Error (AAMs 0.316) | feature descriptors, AAMs, Gabor wavelets and support vector regression (SVR) are used for classification |
|                     |                   | Mean Error (Gabor 0.349) |                                                                            |
|                     |                   | MAEs (AAMs 0.324) |                                                                            |
|                     |                   | MAEs (Gabor0.651) |                                                                            |

As shown in Tables 3 and 4, deep-learning-based approaches outperform conventional approaches.

Table 4. Comparisons of deep-learning-based approaches for FER.

| References          | Datasets          | Accuracy (%) | Methods                                                                 |
|---------------------|-------------------|--------------|-------------------------------------------------------------------------|
| Li et al. [65]      | CK+               | 97.38        | Convolutional neural network (CNN)                                       |
|                     | JAFFE             | 0.14 SD      |                                                                          |
|                     |                   | 97.18        |                                                                          |
|                     |                   | 0.30 SD      |                                                                          |
| Li et al. [55]      | SMIC              | 55.49        | 3D flow-based CNN model for video-based micro-expression recognition     |
|                     | CASME             | 54.44        |                                                                          |
|                     | CASME II          | 59.11        |                                                                          |
| Xie et al. [66]     | CK+               | With 10-fold cross-validation accuracy 95.88 99.32 | Expressional Region Descriptor (SERD) and Multi-Path Variation-Suppressing Network (MPVS-Net) |
|                     | JAFFE             | With 10-fold cross-validation accuracy 95.88 99.32 |                                                                          |
| Li et al. [55]      | RAF-DB AffectNet  | 91.64        | Convolutional neural network (CNN) with attention mechanism (ACNN)       |
|                     | SFEW              |              |                                                                          |
|                     | CK+               | Average of Cclass (96.76) |                                                                            |
|                     |                   | Average of Cbin (98.92) |                                                                            |
|                     |                   | Average of Cclass (72.89) |                                                                            |
|                     |                   | Average of Cbin (90.96) |                                                                            |
|                     |                   | Average of Cclass (53.44) |                                                                            |
|                     |                   | Average of Cbin (84.48) |                                                                            |
| Lopes et al. [56]   | CK+               | Average of Cclass (96.76) | Convolutional neural network and specific image preprocessing steps      |
|                     | JAFFE             | Average of Cbin (98.92) |                                                                          |
|                     | BU-3DFE           | Average of Cclass (72.89) |                                                                          |
|                     |                   | Average of Cbin (90.96) |                                                                          |
|                     |                   | Average of Cclass (53.44) |                                                                          |
|                     |                   | Average of Cbin (84.48) |                                                                          |
| Breuer and Kimmel [57] | CK+ FER2013      | 98.62 ± 0.11 | Convolutional neural network (CNN) with transfer learning               |
|                     |                   | 72 ± 10.5    |                                                                          |
Table 4. Cont.

| References          | Datasets          | Accuracy (%)                   | Methods                                                                 |
|---------------------|-------------------|--------------------------------|-------------------------------------------------------------------------|
| Chu et al. [58]     | BP4D              | Accuracy on 3-fold (53.2)      | Extracted spatial representations by CNN and for temporal dependencies,  |
|                     |                   | Accuracy on 10-fold (82.5)     | LSTMs were used                                                         |
| Hasani and Mahoor [58] | CK+              | 89.5                           | Facial landmark points fed to CNN                                       |
|                     | DISFA             | 56.0                           |                                                                         |
| Jain et al. [61]    | CK+               | 96.17                          | LSTM CNN using texture patterns for facial expression prediction        |
|                     | MMI               | 98.72                          |                                                                         |
| (Zhang et al. 2017) | CK+              | 97.78                          | Part-Based Hierarchical Bidirectional Recurrent Neural Network (PHRNN)   |
|                     | Oulu-CASIA        | 86.25                          | and Multi-Signal Convolutional Neural Network (MSCNN)                   |
|                     | MMI               | 79.29                          |                                                                         |
| Al-Shabi et al. [62]| CK+              | 99.1                           | Dense SIFT and regular SIFT are merged with CNN features               |
|                     | FER-2013          | 73.4                           |                                                                         |
| Jung et al. [63]    | CK+              | 97.25                          | Fusion of deep temporal appearance network (DTAN) and deep temporal     |
|                     | Oulu-CASIA        | 81.46                          | geometry network (DTGN) is employed to predict facial expressions      |
| Yu and Zhang [64]   | SFEW              | Validation accuracy 55.96      | Assembly of DCNNs with seven convolution layers                        |
|                     |                   | Test accuracy 61.29            |                                                                         |

In experimental tests, DL-based FER methods have achieved high precision; however, a range of issues remain that require further investigation:

- As the framework becomes increasingly deeper for preparation, a large-scale dataset and significant computational resources are needed.
- Significant quantities of datasets that are manually compiled and labeled are required.
- A significant amount of memory is required for experiments and testing, which is time-consuming.

The approaches mentioned above, especially those based on deep learning, require massive computation power. Moreover, these approaches are developed for specific emotions, and thus are not suitable to classify other emotional states. Therefore, developing a new framework to be applied to the whole spectrum of emotions would be of significant relevance, and could be expanded to classify complex facial expressions.

7. Conclusions and Future Work

In this paper, a detailed analysis and comparison are presented on FER approaches. We categorized these approaches into two major groups: (1) conventional ML-based approaches and (2) DL-based approaches. The convention ML approach consists of face detection, feature extraction from detected faces and emotion classification based on extracted features. Several classification schemes are used in conventional ML for FER, consisting of random forest, AdaBoost, KNN and SVM. In contrast with DL-based FER methods, the dependency on face physics-based models is highly reduced. In addition, they reduce the preprocessing time to enable “end-to-end” learning in the input images. However, these methods consume more time in both the training and testing phases. Although a hybrid architecture demonstrates better performance, micro-expressions remain difficult tasks to solve due to other possible movements of the face that occur unwillingly.

Additionally, different datasets related to FER are elaborated for the new researchers in this area. For example, human facial emotions have been examined in a traditional database with 2D video sequences or 2D images. However, facial emotion recognition based on 2D data is unable to handle large variations in pose and subtle facial behaviors.
Therefore, recently, 3D facial emotion datasets have been considered to provide better results. Moreover, different FER approaches and standard evaluation metrics have been used for comparison purposes, e.g., accuracy, precision, recall, etc.

FER performance has increased due to the combination of DL approaches. In this modern age, the production of sensible machines is very significant, recognizing the facial emotions of different individuals and performing actions accordingly. It has been suggested that emotion-oriented DL approaches can be designed and fused with IoT sensors. In this case, it is predicted that this will increase FER’s performance to the same level as human beings, which will be very helpful in healthcare, investigation, security and surveillance.

**Funding:** This research received no external funding.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** Not applicable.

**Acknowledgments:** This research was supported by the Artificial Intelligence & Data Analytics Lab (AIDA), CCIS Prince Sultan University, Riyadh, Saudi Arabia.

**Conflicts of Interest:** The authors declare no conflict of interest.

**References**

1. Jabeen, S.; Mehmood, Z.; Mahmood, T.; Saba, T.; Rehman, A.; Mahmood, M.T. An effective content-based image retrieval technique for image visual representations based on the bag-of-visual-words model. *PLoS ONE* **2018**, *13*, e0194526. [CrossRef] [PubMed]

2. Moret-Tatay, C.; Wester, A.G.; Gamermann, D. To Google or not: Differences on how online searches predict names and faces. *Mathematics* **2020**, *8*, 1964. [CrossRef]

3. Ubaid, M.T.; Khalil, M.; Khan, M.U.G.; Saba, T.; Rehman, A. Beard and Hair Detection, Segmentation and Changing Color Using Mask R-CNN. In Proceedings of the International Conference on Information Technology and Applications, Dubai, United Arab Emirates, 13–14 November 2021; Springer: Singapore, 2022; pp. 63–73.

4. Meethongjan, K.; Dzulkifli MRehman, A.; Altameem, A.; Saba, T. An intelligent fused approach for face recognition. *J. Intell. Syst.* **2013**, *22*, 197–212. [CrossRef]

5. Elarbi-Boudihir, M.; Rehman, A.; Saba, T. Video motion perception using optimized Gabor filter. *Int. J. Phys. Sci.* **2011**, *6*, 2799–2806.

6. Joudaki, S.; Rehman, A. Dynamic hand gesture recognition of sign language using geometric features learning. *Int. J. Comput. Vis. Robot.* **2022**, *12*, 1–16. [CrossRef]

7. Abunadi, I.; Albraikan, A.A.; Alzahranî, J.S.; Eltahir, M.M.; Hilal, A.M.; Eldesouki, M.I.; Motvwakel, A.; Yaseen, I. An Automated Glowworm Swarm Optimization with an Inception-Based Deep Convolutional Neural Network for COVID-19 Diagnosis and Classification. *Healthcare* **2022**, *10*, 697. [CrossRef]

8. Yar, H.; Hussain, T.; Khan, Z.A.; Koundal, D.; Lee, M.Y.; Baik, S.W. Vision sensor-based real-time fire detection in resource-constrained IoT environments. *Comput. Intell. Neurosci.* **2021**, *2021*, 5195508. [CrossRef]

9. Yasin, M.; Cheema, A.R.; Kausar, F. Analysis of Internet Download Manager for collection of digital forensic artefacts. *Digit. Investig.* **2010**, *7*, 90–94. [CrossRef]

10. Rehman, A.; Alqahtani, S.; Altameem, A.; Saba, T. Virtual machine security challenges: Case studies. *Int. J. Mach. Learn. Cybern.* **2014**, *5*, 729–742. [CrossRef]

11. Afza, F.; Khan, M.A.; Sharif, M.; Kadry, S.; Manogaran, G.; Saba, T.; Ashraf, I.; Damaševićius, R. A framework of human action recognition using length control features fusion and weighted entropy-variances based feature selection. *Image Vis. Comput.* **2021**, *106*, 104090. [CrossRef]

12. Rehman, A.; Khan, M.A.; Mehmood, Z.; Tariq, U.; Ayesha, N. Microscopic brain tumor detection and classification using 3D CNN and feature selection architecture. *Microsc. Res. Tech.* **2021**, *84*, 133–149. [CrossRef]

13. Haji, M.S.; Alkawaz, M.H.; Rehman, A.; Saba, T. Content-based image retrieval: A deep look at features prospectus. *Int. J. Comput. Vis. Robot.* **2019**, *9*, 14–38. [CrossRef]

14. Alkawaz, M.H.; Mohamad, D.; Rehman, A.; Basori, A.H. Facial animations: Future research directions & challenges. *3D Res.* **2014**, *5*, 12.

15. Saleem, S.; Khan, M.; Ghani, U.; Saba, T.; Abunadi, I.; Rehman, A.; Bahaj, S.A. Efficient facial recognition authentication using edge and density variant sketch generator. *CMC-Comput. Mater. Contin.* **2022**, *70*, 505–521. [CrossRef]

16. Rahim, M.S.M.; Rad, A.E.; Rehman, A.; Altameem, A. Extreme facial expressions classification based on reality parameters. *3D Res.* **2014**, *5*, 22. [CrossRef]
47. Saba, T.; Haseeb, K.; Ahmed, I.; Rehman, A. Secure and energy-efficient framework using Internet of Medical Things for e-healthcare. J. Infect. Public Health 2020, 13, 1567–1575. [CrossRef]
48. Saba, T.; Bokhari ST, F.; Sharif, M.; Yasmin, M.; Raza, M. Fundus image classification methods for the detection of glaucoma: A review. Micros. Res. Tech. 2018, 81, 1105–1121. [CrossRef]
49. Saba, T. Automated lung nodule detection and classification based on multiple classifiers voting. Micros. Res. Tech. 2019, 8, 1601–1609. [CrossRef]
50. Sadad, T.; Munir, A.; Saba, T.; Hussain, A. Fuzzy C-means and region growing based classification of tumor from mammograms using hybrid texture feature. J. Comput. Sci. 2018, 29, 34–45. [CrossRef]
51. Amin, J.; Sharif, M.; Yasmin, M.; Saba, T.; Anjum, M.A.; Fernandes, S.L. A new approach for brain tumor segmentation and classification based on score level fusion using transfer learning. J. Med. Syst. 2019, 43, 326. [CrossRef]
52. Javed, R.; Rahim, M.S.M.; Saba, T.; Rehman, A. A comparative study of features selection for skin lesion detection from dermoscopic images. Netw. Modeling Anal. Health Inform. Bioinform. 2020, 9, 4. [CrossRef]
53. Li, J.; Wang, Y.; See, J.; Liu, W. Micro-expression recognition based on 3D flow convolutional neural network. Pattern Anal. Appl. 2019, 22, 1331–1339. [CrossRef]
54. Li, B.Y.; Mian, A.S.; Liu, W.; Krishna, A. Using kinect for face recognition under varying poses, expressions, illumination and disguise. In Proceedings of the IEEE workshop on applications of computer vision (WACV), Clearwater Beach, FL, USA, 15–17 January 2013; pp. 186–192.
55. Li, Y.; Zeng, J.; Shan, S.; Chen, X. Occlusion aware facial expression recognition using CNN with attention mechanism. IEEE Trans. Image Processing 2018, 28, 2439–2450. [CrossRef]
56. Lopes, A.T.; De Aguiar, E.; De Souza, A.F.; Oliveira-Santos, T. Facial expression recognition with Convolutional Neural Networks: Coping with few data and the training sample order. Pattern Recognit. 2017, 61, 610–628. [CrossRef]
57. Breuer, R.; Kimmel, R. A deep learning perspective on the origin of facial expressions. arXiv 2017, arXiv:1705.01842.
58. Chu, W.S.; De La Torre, F.; Cohn, J.F. Learning spatial and temporal cues for multi-label facial action unit detection. In Proceedings of the 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017), Washington, DC, USA, 30 May–3 June 2017; pp. 25–32.
59. Hasani, B.; Mahoor, M.H. Facial expression recognition using enhanced deep 3D convolutional neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Honolulu, HI, USA, 21–26 July 2017; pp. 30–40.
60. Zhang, K.; Huang, Y.; Du, Y.; Wang, L. Facial expression recognition based on deep evolutionary spatial-temporal networks. IEEE Trans. Image Processing 2017, 26, 4193–4203. [CrossRef]
61. Jain, D.K.; Zhang ZHuang, K. Multi angle optimal pattern-based deep learning for automatic facial expression recognition. Pattern Recognit. Lett. 2020, 139, 157–165. [CrossRef]
62. Al-Shabi, M.; Cheah, W.P.; Connie, T. Facial Expression Recognition Using a Hybrid CNN-SIFT Aggregator. arXiv 2016, arXiv:1608.02833.
63. Jung, H.; Lee, S.; Yim, J.; Park, S.; Kim, J. Joint fine-tuning in deep neural networks for facial expression recognition. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 2983–2991.
64. Yu, Z.; Zhang, C. Image based static facial expression recognition with multiple deep network learning. In Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, Seattle, WA, USA, 9–13 November 2015; pp. 435–442.
65. Li, K.; Jin, Y.; Akram, M.W.; Han, R.; Chen, J. Facial expression recognition with convolutional neural networks via a new face cropping and rotation strategy. Vis. Comput. 2020, 36, 391–404. [CrossRef]
66. Xie, S.; Hu, H.; Wu, Y. Deep multi-path convolutional neural network joint with salient region attention for facial expression recognition. Pattern Recognit. 2019, 92, 177–191. [CrossRef]
67. Nasir, I.M.; Raza, M.; Shah, J.H.; Khan, M.A.; Rehman, A. Human action recognition using machine learning in uncontrolled environment. In Proceedings of the 2021 1st International Conference on Artificial Intelligence and Data Analytics (CAIDA), Riyadh, Saudi Arabia, 6–7 April 2021; pp. 182–187.
68. Rehman, A.; Haseeb, K.; Saba, T.; Lloret, J.; Tariq, U. Secured Big Data Analytics for Decision-Oriented Medical System Using Internet of Things. Electronics 2021, 10, 1273. [CrossRef]
69. Harouni, M.; Rahim, M.S.M.; Al-Rodhaan, M.; Saba, T.; Rehman, A.; Al-Dhelaan, A. Online Persian/Arabic script classification without contextual information. Imaging Sci. J. 2014, 62, 437–448. [CrossRef]
70. Itikhar, S.; Fatima, K.; Rehman, A.; Almazayd, A.S.; Saba, T. An evolution based hybrid approach for heart diseases classification and associated risk factors identification. Biomed. Res. 2017, 28, 3451–3455.
71. Lucey, P.; Cohn, J.F.; Kanade, T.; Saragih, J.; Ambadar, Z.; Matthews, I. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In Proceedings of the IEEE computer society conference on computer vision and pattern recognition-workshops, San Francisco, CA, USA, 13–18 June 2010; pp. 94–101.
72. Jamal, A.; Hazim Alkawaz, M.; Rehman, A.; Saba, T. Retinal imaging analysis based on vessel detection. Micros. Res. Tech. 2017, 80, 799–811. [CrossRef]
73. Neamah, K.; Mohamed, D.; Saba, T.; Rehman, A. Discriminative features mining for offline handwritten signature verification. 3D Res. 2014, 5, 2. [CrossRef]
74. Ramzan, F.; Khan, M.U.G.; Iqbal, S.; Saba, T.; Rehman, A. Volumetric segmentation of brain regions from MRI scans using 3D convolutional neural networks. IEEE Access 2020, 8, 103697–103709. [CrossRef]
75. Keltner, D.; Sauter, D.; Tracy, J.; Cowen, A. Emotional expression: Advances in basic emotion theory. *J. Nonverbal Behav.* 2019, 43, 133–160. [CrossRef]

76. Phetchanchai, C.; Selamat, A.; Rehman, A.; Saba, T. Index financial time series based on zigzag-perceptually important points. *J. Comput. Sci.* 2010, 6, 1389–1395.

77. Saba, T.; Rehman, A.; Altameem, A.; Uddin, M. Annotated comparisons of proposed preprocessing techniques for script recognition. *Neural Comput. Appl.* 2014, 25, 1337–1347. [CrossRef]

78. Saba, T.; Altameem, A. Analysis of vision based systems to detect real time goal events in soccer videos. *Appl. Artif. Intell.* 2013, 27, 656–667. [CrossRef]

79. Ullah, H.; Saba, T.; Islam, N.; Abbas, N.; Rehman, A.; Mehmood, Z.; Anjum, A. An ensemble classification of exudates in color fundus images using an evolutionary algorithm based optimal features selection. *Microsc. Res. Tech.* 2019, 82, 361–372. [CrossRef] [PubMed]

80. Sharif, U.; Mehmood, Z.; Mahmood, T.; Javid, M.A.; Rehman, A.; Saba, T. Scene analysis and search using local features and support vector machine for effective content-based image retrieval. *Artif. Intel. Rev.* 2019, 52, 901–925. [CrossRef]

81. Sharif, U.; Mehmood, Z.; Mahmood, T.; Javid, M.A.; Rehman, A.; Saba, T. Scene analysis and search using local features and support vector machine for effective content-based image retrieval. *Artif. Intel. Rev.* 2019, 52, 901–925. [CrossRef] [PubMed]

82. Sharif, U.; Mehmood, Z.; Mahmood, T.; Javid, M.A.; Rehman, A.; Saba, T. Scene analysis and search using local features and support vector machine for effective content-based image retrieval. *Artif. Intel. Rev.* 2019, 52, 901–925. [CrossRef] [PubMed]

83. Sharif, U.; Mehmood, Z.; Mahmood, T.; Javid, M.A.; Rehman, A.; Saba, T. Scene analysis and search using local features and support vector machine for effective content-based image retrieval. *Artif. Intel. Rev.* 2019, 52, 901–925. [CrossRef] [PubMed]

84. Sharif, U.; Mehmood, Z.; Mahmood, T.; Javid, M.A.; Rehman, A.; Saba, T. Scene analysis and search using local features and support vector machine for effective content-based image retrieval. *Artif. Intel. Rev.* 2019, 52, 901–925. [CrossRef] [PubMed]

85. Sharif, U.; Mehmood, Z.; Mahmood, T.; Javid, M.A.; Rehman, A.; Saba, T. Scene analysis and search using local features and support vector machine for effective content-based image retrieval. *Artif. Intel. Rev.* 2019, 52, 901–925. [CrossRef] [PubMed]

86. Sharif, U.; Mehmood, Z.; Mahmood, T.; Javid, M.A.; Rehman, A.; Saba, T. Scene analysis and search using local features and support vector machine for effective content-based image retrieval. *Artif. Intel. Rev.* 2019, 52, 901–925. [CrossRef] [PubMed]

87. Sharif, U.; Mehmood, Z.; Mahmood, T.; Javid, M.A.; Rehman, A.; Saba, T. Scene analysis and search using local features and support vector machine for effective content-based image retrieval. *Artif. Intel. Rev.* 2019, 52, 901–925. [CrossRef] [PubMed]

88. Sharif, U.; Mehmood, Z.; Mahmood, T.; Javid, M.A.; Rehman, A.; Saba, T. Scene analysis and search using local features and support vector machine for effective content-based image retrieval. *Artif. Intel. Rev.* 2019, 52, 901–925. [CrossRef] [PubMed]

89. Sharif, U.; Mehmood, Z.; Mahmood, T.; Javid, M.A.; Rehman, A.; Saba, T. Scene analysis and search using local features and support vector machine for effective content-based image retrieval. *Artif. Intel. Rev.* 2019, 52, 901–925. [CrossRef] [PubMed]

90. Sharif, U.; Mehmood, Z.; Mahmood, T.; Javid, M.A.; Rehman, A.; Saba, T. Scene analysis and search using local features and support vector machine for effective content-based image retrieval. *Artif. Intel. Rev.* 2019, 52, 901–925. [CrossRef] [PubMed]

91. Sharif, U.; Mehmood, Z.; Mahmood, T.; Javid, M.A.; Rehman, A.; Saba, T. Scene analysis and search using local features and support vector machine for effective content-based image retrieval. *Artif. Intel. Rev.* 2019, 52, 901–925. [CrossRef] [PubMed]

92. Sharif, U.; Mehmood, Z.; Mahmood, T.; Javid, M.A.; Rehman, A.; Saba, T. Scene analysis and search using local features and support vector machine for effective content-based image retrieval. *Artif. Intel. Rev.* 2019, 52, 901–925. [CrossRef] [PubMed]

93. Sharif, U.; Mehmood, Z.; Mahmood, T.; Javid, M.A.; Rehman, A.; Saba, T. Scene analysis and search using local features and support vector machine for effective content-based image retrieval. *Artif. Intel. Rev.* 2019, 52, 901–925. [CrossRef] [PubMed]

94. Sharif, U.; Mehmood, Z.; Mahmood, T.; Javid, M.A.; Rehman, A.; Saba, T. Scene analysis and search using local features and support vector machine for effective content-based image retrieval. *Artif. Intel. Rev.* 2019, 52, 901–925. [CrossRef] [PubMed]

95. Sharif, U.; Mehmood, Z.; Mahmood, T.; Javid, M.A.; Rehman, A.; Saba, T. Scene analysis and search using local features and support vector machine for effective content-based image retrieval. *Artif. Intel. Rev.* 2019, 52, 901–925. [CrossRef] [PubMed]

96. Sharif, U.; Mehmood, Z.; Mahmood, T.; Javid, M.A.; Rehman, A.; Saba, T. Scene analysis and search using local features and support vector machine for effective content-based image retrieval. *Artif. Intel. Rev.* 2019, 52, 901–925. [CrossRef] [PubMed]

97. Sharif, U.; Mehmood, Z.; Mahmood, T.; Javid, M.A.; Rehman, A.; Saba, T. Scene analysis and search using local features and support vector machine for effective content-based image retrieval. *Artif. Intel. Rev.* 2019, 52, 901–925. [CrossRef] [PubMed]

98. Sharif, U.; Mehmood, Z.; Mahmood, T.; Javid, M.A.; Rehman, A.; Saba, T. Scene analysis and search using local features and support vector machine for effective content-based image retrieval. *Artif. Intel. Rev.* 2019, 52, 901–925. [CrossRef] [PubMed]

99. Sharif, U.; Mehmood, Z.; Mahmood, T.; Javid, M.A.; Rehman, A.; Saba, T. Scene analysis and search using local features and support vector machine for effective content-based image retrieval. *Artif. Intel. Rev.* 2019, 52, 901–925. [CrossRef] [PubMed]

100. Sharif, U.; Mehmood, Z.; Mahmood, T.; Javid, M.A.; Rehman, A.; Saba, T. Scene analysis and search using local features and support vector machine for effective content-based image retrieval. *Artif. Intel. Rev.* 2019, 52, 901–925. [CrossRef] [PubMed]