Hindawi
Human Behavior and Emerging Technologies
Volume 2023, Article ID 9790005, 27 pages
https://doi.org/10.1155/2023/9790005

*Review Article*

# Academic Emotion Classification Using FER: A Systematic Review

**Jeniffer Xin-Ying Lek and Jason Teo**

*Faculty of Computing and Informatics, Universiti Malaysia Sabah, Jalan UMS, 88400 Kota Kinabalu, Sabah, Malaysia*

Correspondence should be addressed to Jason Teo; jtwteo@ums.edu.my

Received 7 September 2022; Revised 21 February 2023; Accepted 25 April 2023; Published 3 May 2023

Academic Editor: Sarah E. Domoff



Facial emotion expressions are among the most potent, natural, and powerful means of human communication. Due to the COVID-19 pandemic, educational institutions worldwide are forced to switch rapidly to remote and online learning. Students are currently in an emergency state and must adapt to various and readily accessible learning methods, such as mobile learning applications or an e-learning system. A systematic literature review (SLR) is conducted to extract and synthesize information such as the emotion classifier used in the facial expression recognition (FER) system, the dataset used, the preprocessing technique applied, the feature extraction approach used, and the strength and limitation of the previous studies. Based on the search criteria, 701 publications were initially retrieved from five different digital databases, of which 48 studies have been chosen as primary studies for further analysis. Based on the findings of this study, the deep learning approach is the most frequently adopted approach in classifying student emotions during online learning. FER-2013 is the most commonly used FER dataset in FER studies, while DAiSEE is the most used academic emotion dataset. Moreover, support vector machine (SVM) is the conventional learning emotion classifier that is widely used in the FER systems, while convolutional neural network (CNN) is the most frequently used deep learning classifier. Next, it was found that the number of real-time FER systems is less than that of non-real-time FER systems. Finally, the top-1 accuracy of 94.6% was achieved by the long-term recurrent convolutional network on the academic emotion dataset, and the limitation is that it has low illumination and a lack of frontal pose.

## 1. Introduction

Facial emotion expressions are generally considered to be a powerful means of communication among humans. However, research has shown that cultural and individual differences can exist in how people interpret and respond to them [1]. Hence, it is important to address automatic emotion recognition based solely on facial emotion expressions with caution since it may not account for the diverse and subtle ways in which emotions are expressed and interpreted by individuals and across cultures.

Online learning can be defined as learning that takes place over the internet at the learner's own pace or in real time, depending on the platform used [2]. According to a global survey, 85% of universities and other educational institutions adopted online learning as their primary teaching mode during the COVID-19 pandemic [3]. Students are currently in an emergency state and must adapt to various and readily accessible learning methods, such as mobile learning applications or an e-learning system. Distance learning and e-learning are not novel concepts for learners. Nevertheless, the pandemic has highlighted the importance of exploring online teaching and learning opportunities [4]. This situation has demonstrated the pros and cons of educational systems when confronted with the challenges of digitalization.

According to the study by Krithika and Priya, emotion plays a crucial role in online learning as it can impact the student's interest in lectures [5]. An enthusiastic student has a higher chance of successful learning compared to a bored student, making it crucial to adapt learning to students' emotions [6]. Furthermore, students may experience a variety of emotional or mental states in online learning, which can impact their learning process. During the pandemic, many students suffered from depression, academic stress, or anxiety in e-learning because they struggled to

adapt to the new norm. Therefore, an emotion recognition system could assist educators in perceiving their students' emotional state in online learning.

Facial emotions are among the most potent, natural, and universal messages for individuals to convey their emotions and thoughts irrespective of gender, ethnicity, and nationality [7]. Emotions are described as physical and mental states that help people deal with various circumstances. An individual is said to be attentive when his mind is focused on a certain topic. Educators interpret attention as a mental state in which learners focus on something since it is a prerequisite for learning and motivation in a classroom.

According to a meta-analysis by Camacho-Morles et al., the enjoyment of learning is positively correlated with academic performance, while negative emotions, such as anger and boredom, are negatively correlated with academic performance [8]. Students' emotions and motivation affect their learning ability, whereas positive emotions and moods can sometimes lead to more precise decision-making, creativity, and adaptability [9–11]. Hence, individuals must be prepared on all levels, including emotionally, physically, socially, and mentally, to effectively absorb the information during learning. Students who fail to understand and process information while learning may experience emotions like confusion, fatigue, boredom, or exhaustion [9]. This can lead to a change in student behaviour, leading them to skip the online lectures or drop out of the course. Consequently, students struggling with unstable or negative emotions can find it difficult to learn effectively through online learning.

Identifying the emotion of students can bring many benefits, such as determining which type of teaching style or teaching materials have the potential to boost students' positive feelings. Hence, the lecture materials or teaching style can be tailored to make students interested in every online lecture. Rothkrantz stated that positive emotions in students could favour the intention to engaged in online learning [12]. In contrast, negative emotions can lead to negative learning outcomes, influence students' educational path, and cause them to lose interest in learning. An enthusiastic student had a higher chance of successful learning compared to a bored student. Hence, learning must adapt to the emotion of the students [6].

A systematic literature review (SLR) is conducted to obtain an overview of what has been done on the student FER system in the education field. SLR indicates the potential research gaps in a specific problem area and provides guidance to researchers and practitioners that are interested in carrying out new studies on the particular problem area. All associated research papers are retrieved from various digital sources, integrated, and discussed to answer the research questions stated. The SLR study yields new insights and assists new researchers in learning better about the state-of-the-art.

In this paper, the sections are structured as follows. Following the introduction is the background section, which details the facts regarding the problems of FER in online learning and the taxonomy of FER, which is split into preprocessing, feature extraction, and emotion recognition. After that, the methodology for identifying the relevant research studies will be discussed in the methodology section. The result section will summarize the results of completing the SLR phases, while the approaches used in FER in online learning will be reviewed and analyzed in the discussion section. Then, the research challenges and limitations will be discussed. Last but not least, this review paper will be concluded in the conclusion section.

## 2. Background

In this section, the background and significance of facial emotion recognition (FER) in online learning will be included. Firstly, the online learning problem will be discussed, followed by the significance of FER, the main stages of FER, and finally, the taxonomy of FER.

*2.1. Online Learning Problem.* Distance and self-paced learning methods appear to have shattered the bonds of friendship and social interaction between students in the classroom. Thus, video conferencing platforms like Google Meet, Webex, and Microsoft Teams have facilitated real-time interaction between students and educators in online learning. Nevertheless, educators are unable to determine the students' interest level in online learning via video conferencing platforms [13]. This restriction can be attributed to a number of different reasons. Firstly, there is a possibility that interactive learning may not be supported by online applications. Students' privacy was safeguarded by features like microphone muting, shuttering, and limiting the camera or webcam recording capabilities during the online class via these platforms. As a result, most students prefer to use these provided features in online classes. Another drawback is that these tools have a narrow field of view, which causes educators can only observe the students' faces but not their postures and surroundings. Consequently, educators are not able to fully grasp the true emotions of their students as they can in traditional classroom settings. Besides that, students use a variety of gadgets, including tablets, smartphones, and computers, all of which have the potential to degrade visual quality.

There is a critical lack of academic engagement in large-scale online education like massive open online courses (MOOCs), which are provided by many organizations, including some of the world's best-known institutions with a wide range of offerings and have millions of students enrolled [14]. They are able to reach a large scale, but this comes with a significant drawback since the proportionately smaller number of educators cannot accommodate the large number of students they work with. As a result, educators are unable to monitor and participate in their students' progress and engagement, which frequently results in a waterfall-style, one-way transfer of knowledge from educators to students.

*2.2. Facial Emotion Recognition (FER).* In human communication, facial emotion is crucial to assist humans in understanding people's intentions. The fluency, precision, and truthfulness of the interaction or communication can be

enhanced by facial emotion recognition. This method of recognition is functional when it comes to deciphering the interactions between humans and computers.

Previous studies have shown that two-thirds of interpersonal communication is conveyed by nonverbal elements. Facial emotions are part of crucial sources of information in human communication among the nonverbal elements [15]. The ability to recognize facial emotions is fundamental in order to effective interpersonal communication. In fact, emotion recognition is crucial to the experience of empathy, prosocial behaviour prediction, and the ability model of emotional intelligence. It is a difficult task because of some problems, such as action similarities, and large head positions. The study by Sariyanidi et al. revealed that various components contribute to the effectiveness of FER approaches, including factors such as precise face registration, effective representation methods, and accurate emotion recognition algorithms [16].

There are several approaches to recognizing an individual's emotions, and the acquisition of facial-based features is the most commonly used approach [17]. The facial channel is universally recognized as the dominant channel of emotional expression in humans among automatic emotion recognition techniques, resulting in facial emotion recognition being the most researched among the various channels of emotional expression.

A number of approaches can be applied to assess the student's emotions in online learning, such as through students' facial expressions, eye movements, gestures, and posture, as well as feedback checklists from students [18]. Nonetheless, the educator can only view the students' faces using the currently available technologies for online classes [19]. Besides that, sensors can be utilized for FER inputs, such as a camera, electroencephalograph (EEG), electrocardiogram (ECG), and electromyography (EMG). However, the camera is a promising sensor since it does not have to be worn and gives FER the most detailed indication [15].

*2.3. Main Stages of FER.* In this section, the main stages of FER will be discussed. The process typically involves three stages in FER: preprocessing, feature extraction, and emotion recognition. These stages are all essential for building an accurate and effective FER system.

*2.3.1. Preprocessing.* Preprocessing is a process that may be taken before the feature extraction in order to enhance the FER system's overall performance [20]. The feature extraction step can be better tailored using preprocessing to reduce noise and redundant data [21].

In FER, face detection, dimension reduction, and normalization are the crucial preprocessing steps before heading to the feature extraction step [22]. Face detection, the first prerequisite phase in FER, involves detecting a face inside a frame or an image and removing any pixels that do not contribute significantly. Face detection is a challenging task since human faces come in a variety of shapes and sizes. Therefore, the face detection algorithm significantly impacts the issue mentioned above. Examples of face detection algorithms include Viola-Jones, genetic algorithms, linear discriminant analysis (LDA), and principal component analysis (PCA).

Next, an approach called dimension reduction is applied to narrow down the number of variables to a set of principal variables [22]. When there are more features to consider, it is more difficult to visualize the training set and perform the necessary steps to improve it. In this context, PCA and LDA are useful algorithms for handling the mentioned issue. Besides that, normalization is a term that is interchangeably used with feature scaling. Following the dimension reduction process, the reduced features are normalized in a manner that does not misrepresent the variations in the range of features' values. In order to speed up the training process and enhance the numerical stability of a model, many different normalization approaches, such as unit vector normalization, min-max normalization, and Z normalization, can be applied.

*2.3.2. Feature Extraction.* The process of extracting the important features, including geometric features, appearance-based features, or physiological features for FER, can result in smaller and more detailed attribute sets. These attribute sets consist of features like the distance between a pair of eyes, the distance between eyes and lips, and the edges, diagonal, and corners of a face, which aid in more rapid learning of previously trained data.

Appearance-based extraction and geometric-based extraction are the two approaches for feature extraction. Features such as corner and edge features may be extracted using the geometric-based extraction method. In order to extract geometric features, the position of the face components must first be recognized and then depicted using a set of feature points, also known as landmarks or contours [23]. Subsequently, the $(x, y)$ coordinates are used to generate a feature vector. The feature vector, which contains geometric information of face components, is computed using the landmark points' distance, the points' arrangement, and the slope of connected lines.

Meanwhile, for the appearance-based extraction technique, salient point features are used in order to maintain the location of the eyes and the form of the lips and eyebrows, as well as other key facial features. This technique does not require face points; instead, it determines the texture information of facial images based on the grey level values of the pixels, along with the connection between each pixel and its set of neighbouring pixels. Furthermore, this technique is often accomplished through a variety of texture descriptors or image filters.

The specific features that are extracted can vary depending on the approach used for FER, and the choice of features can greatly impact the performance of the emotion recognition system. Therefore, feature extraction is an important step in the FER process and requires careful consideration and evaluation.

*2.3.3. Emotion Classification.* The stage after feature extraction in FER is emotion classification, in which the classifier sorts different expressions into the appropriate categories. Various classification algorithms, including conventional

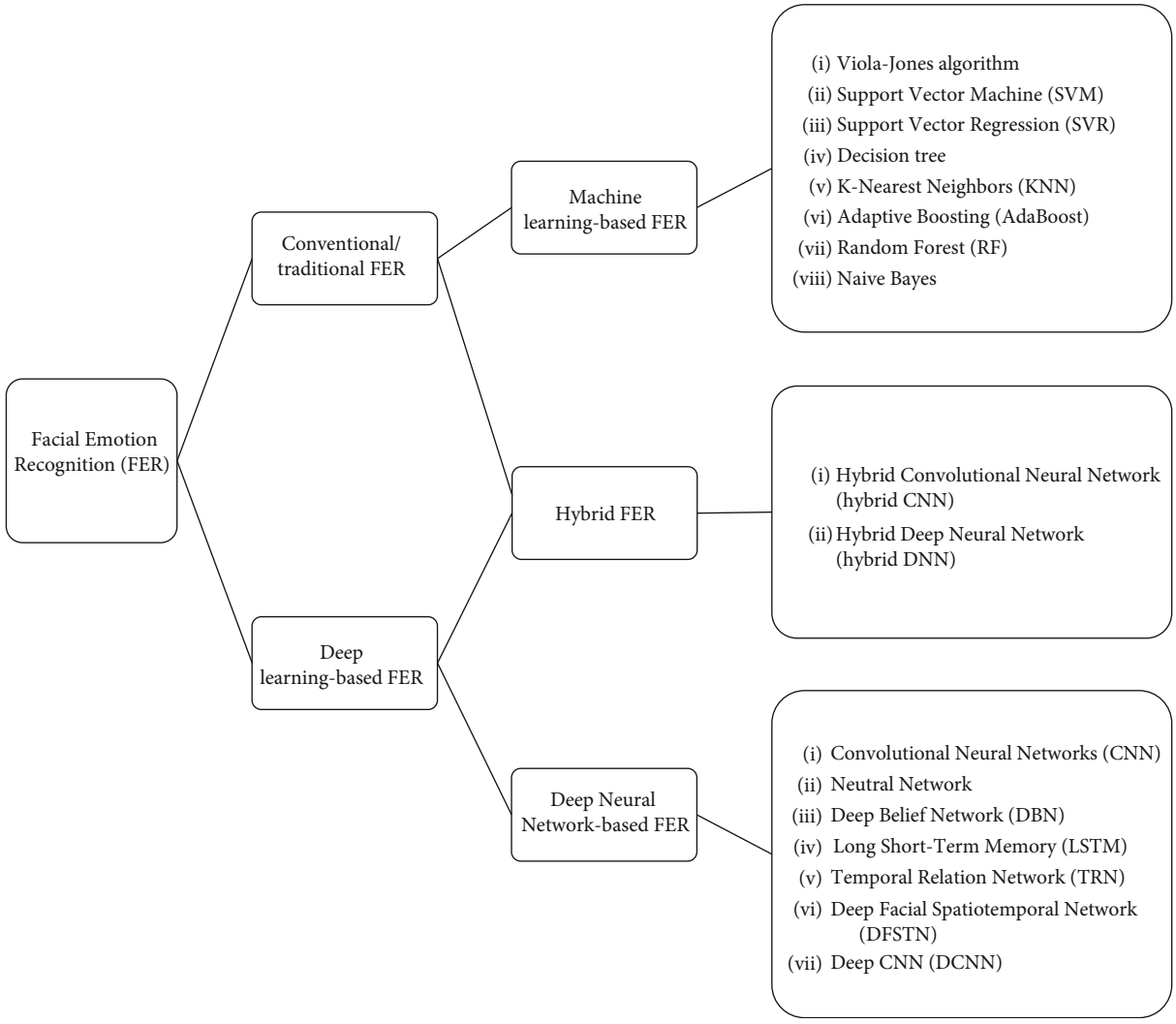


Figure 1: Taxonomy of facial emotion recognition (FER).

learning algorithms and deep learning algorithms, are widely used in emotion classification. CNN is the most widely applied classification algorithm nowadays. The fact that it can be implemented directly on the input image without using any facial detection or feature extraction algorithms makes it the most efficient algorithm [24]. Despite this, it still achieves a higher level of accuracy than the input data.

Human emotions are inconstant since they go through cycles of highs and lows. Thus, classifying emotions based on context is very challenging.

*2.4. Taxonomy of FER.* This section introduces the taxonomy of FER based on the technology used, specifically conventional and deep learning-based approaches. An illustration of the FER comprehensive taxonomy is presented in Figure 1.

*2.4.1. Conventional/Traditional FER.* This approach used handcrafted features extracted from facial emotion expressions, which were then classified using machine learning algorithms [25]. Conventional FER can be further classified as machine learning-based FER. This approach uses machine learning algorithms such as support vector machines (SVM), decision trees, random forests, Naïve Bayes, and K-nearest neighbors to recognize and classify facial expressions.

*2.4.2. Deep Learning-Based FER.* This approach applied deep learning algorithms that allow the automatic extraction of features and classification [26]. Deep learning-based FER can be further classified into hybrid FER and deep neural network-based FER. Hybrid FER combines conventional FER methods like feature extraction and selection with deep learning techniques to improve the model's accuracy [27]. On the other hand, deep neural network-based FER deep neural networks, such as CNNs and RNNs, learn and extract features directly from raw facial images, which are subsequently classified into various categories of emotions.

## 3. Methodology

A review protocol is established before the SLR is performed. The SLR was carried out in accordance with the prominent SLR guidelines that were published in 2007 [28]. A review protocol specifies the approaches used to perform SLR. In

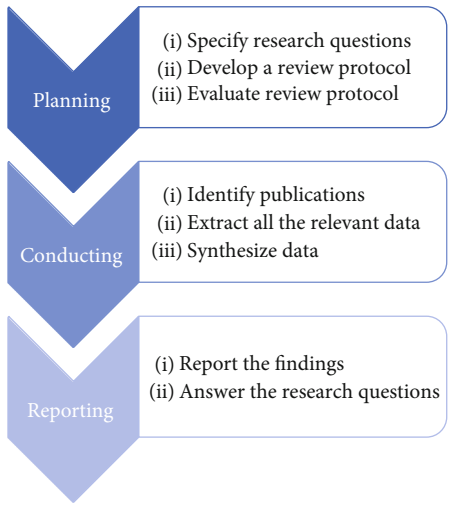


Figure 2: Phase of the systematic literature review.

order to minimize the potential for publication bias, a predefined protocol is required. First of all, the research questions are identified. The digital databases will be used to find relevant research papers when the research questions are ready. Databases such as Science Direct, IEEE Xplore, Springer Link, Google Scholar, and Scopus were used in this SLR. There are three phases in the systematic literature review, including the planning phase, conducting phase, and reporting phase, as presented in Figure 2.

In the first phase, the need for SLR is identified, the research questions are specified, a review protocol is developed, and the review protocol is evaluated. The review protocol evaluation was recursive. The search string and domain list were frequently altered until the search results showed the results of each identified domain found. Next, the publications were identified and chosen by searching the available databases during the conducting phase. The data extraction was done where the authors' details, publication types, publication year, and other details about the research questions were collected. After the proper extraction of all the relevant data, a data synthesis will be made to present an overview of the related studies published up to now. The review was concluded during the final stage by reporting the findings and answering the research questions. The review must be thoroughly reported in adequate detail in order for readers to evaluate the comprehensiveness of the search. The unfiltered search results should be stored in case they need to be analyzed again.

Five research questions are specified to guide this SLR:

RQ1: What is the most frequently adopted approach in classifying student emotion during online learning for the student FER systems in recent years? (Conventional machine learning/deep learning/hybrid)

RQ2: Which dataset is used for the student FER systems in online learning

RQ3: What is the most frequently used emotion classifier in the student FER systems

RQ4: Do the existing student FER systems work in real time

RQ5: What is the accuracy and limitation of previous studies that used the academic emotion dataset

A systematic search was carried out in five digital databases, which are IEEE Xplore, ScienceDirect, Springer Link, Scopus, and Google Scholar, to answer the research questions presented in the previous subsection. The initial search input was "Facial emotion recognition" AND "Online learning". The final search string was as follows: (("facial emotion recognition" OR "facial emotion detection" OR "facial emotion classification) AND ("e-learning" OR "online learning") AND ("deep learning" OR "machine learning")). There were 701 papers initially retrieved following the execution of the stated search string.

Exclusion criteria (EC) were used for study evaluation and assessment to determine the boundaries of the SLR to exclude irrelevant studies. Six ECs are listed as follows:

EC1: publication is a survey or review paper
EC2: publication has been published before 2018
EC3: publication without full text available
EC4: duplicate publication from multiple sources
EC5: publication not written in English
EC6: papers are not computer science-related

Following the application of the listed ECs, only 48 studies have been left for further review. In order to answer the research questions accordingly, the data from the selected publications were extracted and synthesized. Figure 3 illustrates the diagram of the study selection process.

An accurate and effective FER model can assist educators in evaluating the emotion of students in online learning. Various approaches were applied to classify student emotions in both classroom and online learning environments. This review article aims to investigate how machine learning and deep learning are used in student emotion recognition systems based on facial expressions in previous studies.

Review or survey papers are one of the exclusion criteria during the analysis of retrieved publications. The publications that have been omitted are related work that will be addressed in this section. Dewan et al. performed a review study on engagement detection in online learning in 2019 [18]. The paper concluded that the computer vision-based approaches have some constraints, although they are found to be effective in engagement detection. For example, the existing algorithms face difficulties in analyzing facial occlusions and head movements, so the features cannot be extracted from certain video segments, thereby resulting in data loss. In addition, very few available online datasets can be used to detect student emotions in online learning.

Li and Deng published a survey paper about deep facial expression recognition [29]. A detailed review of deep facial expression recognition was presented in their survey, such as algorithms and datasets that clarify fundamental issues such as overfitting due to insufficient expression-unrelated variations and training data. Besides that, the established deep neural networks and associated FER training methods implemented are addressed based on static and dynamic image sequences, along with their pros and cons. Furthermore, the difficulties and opportunities in the FER field and the prospect of developing robust, deep FER systems are also reviewed in their survey paper.

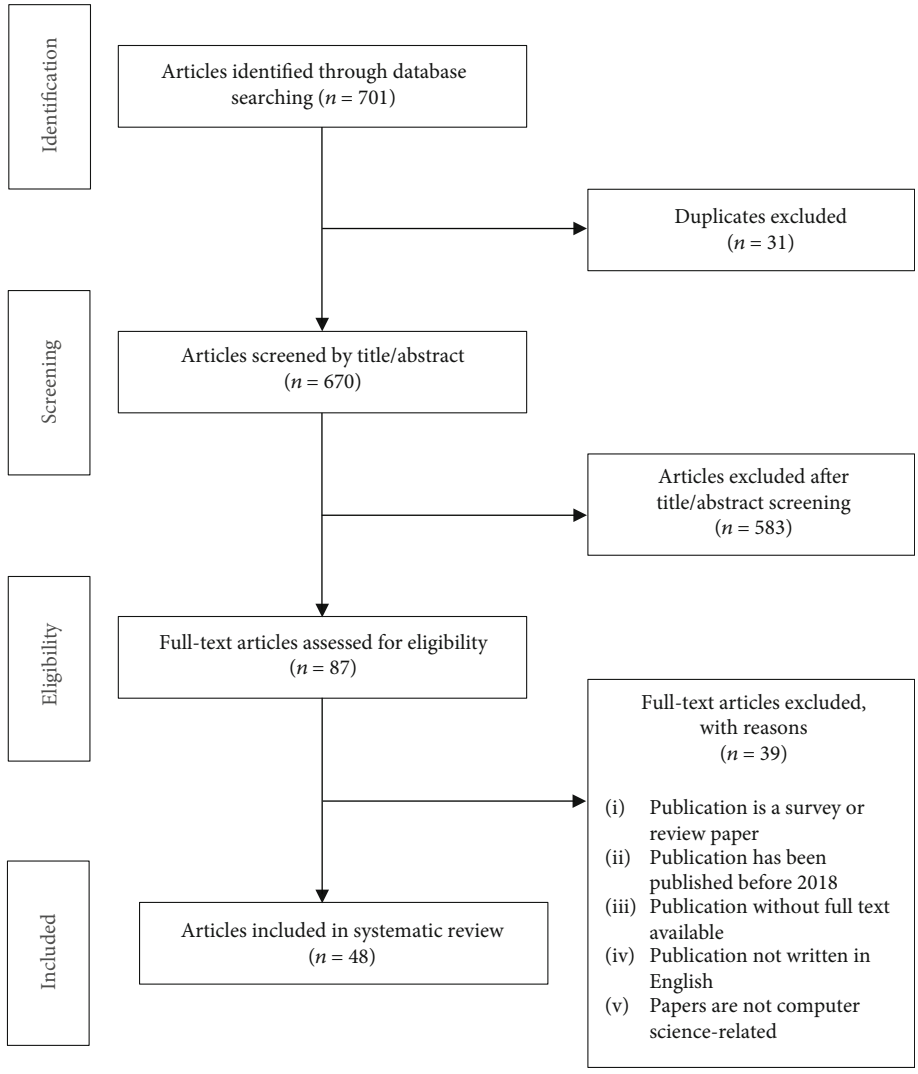


Figure 3: Flow diagram of the study selection process.

Table 1: Overview of search results and publication selection.

| Database | After automated and manual search | After applying exclusion criteria | Percentage of papers (%) |
|---|---|---|---|
| IEEE Xplore | 229 | 19 | 39.58 |
| ScienceDirect | 139 | 2 | 4.17 |
| Google Scholar | 96 | 12 | 25.00 |
| Scopus | 51 | 6 | 12.50 |
| Springer Link | 186 | 9 | 18.75 |
| Total | **701** | **48** | **100.00** |

## 4. Results

In this section, the publications and information related to the FER approach used by the researchers to classify student emotions in online learning will be reviewed and discussed. In addition, the performance of the machine learning algorithms used in the FER will be examined and investigated.

A total of 48 publications were selected to be included in this review paper. Table 1 indicates the number of publications initially retrieved and the number of publications following the application of exclusion criteria. Table 2 summarizes the important information from each publication, such as the emotion classifier and dataset used, the accuracy of the classifier, preprocessing approach, the feature extraction method, strengths, and limitations.

Figure 4 illustrates the year-wise distribution of the primary studies in the previous five years. The statistic presented in Figure 4 indicates the patterns that can be seen

Table 2: Selected retrieved publications.

| Retrieved from | Authors, year | Emotion classifier | Dataset | Preprocessing | Feature extraction | Results | Strength | Limitation |
|---|---|---|---|---|---|---|---|---|
| IEEE Xplore | G. Li and Wang, 2018 [30] | SVM+CNN | FER-2013 | Histogram equalization; frontal_face_detector detector of the Dlib library is loaded for face detection. | Method based on the global feature & geometric features | N/A | Able to analyze the learners' facial images in real time | Focus only on general emotions |
| IEEE Xplore | El Hammoumi et al., 2018 [31] | CNN | CK+ and the KDEF | OpenCV | CNN | Accuracy: 97.18% | Achieved high detection accuracy | Focus only on general emotions |
| IEEE Xplore | Ma et al., 2018 [32] | CNN | FER-2013 | Cascaded Haar feature real-time facial detection algorithm based on AdaBoost | CNN | Identification error score: 0.14 | Real-time detection | Focus only on general emotions |
| ScienceDirect | D. Yang et al., 2018 [33] | Neural network | JAFFE | Haar cascades | Neural network | Accuracy:<br>(i) Sad: 76%<br>(ii) Surprise: 87.72%<br>(iii) Happy: 94%<br>(iv) Anger: 87.66%<br>(v) Disgust: 82.76<br>(vi) Fear: 79.73% | Achieved high classification accuracy for "happy" emotion | (i) Long processing time<br>(ii) Does not involve the illumination and pose of the image |
| IEEE Xplore | Candra Kirana et al., 2018 [34] | Viola-Jones algorithm | Ten student expressions in the class | Viola-Jones algorithm | Viola-Jones algorithm | Accuracy: 74% | Fastest algorithm when compared to the other two algorithms, Viola-Jones+neural networks and neural networks | (i) Small dataset<br>(ii) Has lower detection accuracy compared to more complex algorithms (Viola-Jones +neural networks)<br>(iii) Only works on forward-facing faces |
| IEEE Xplore | Dewan et al., 2018 [35] | DBN | DAiSEE | Viola-Jones algorithm | Local directional pattern (LDP) | Accuracy:<br>(i) Two level: 90.89%<br>(ii) Three level: 87.25% | (i) Achieved high accuracy for two-level engagement detection<br>(ii) Robustness in classification | Unknown direct correlation between engagement and actual task performance |
| Scopus | Sharma and Mansotra, 2019 [36] | CNN | FER-2013 | Viola-Jones algorithm | Haar cascades extraction | N/A | Able to analyze emotions in real time | Focus only on general emotions |
| Scopus | Sharma and Mansotra, 2019 [37] | CNN | FER-2013 and emotional corpus | Viola-Jones algorithm | Haar cascades extraction | N/A | Achieved better classification accuracy | Focus only on general emotions |
| IEEE Xplore | Mao et al., 2019 [38] | SVM | 617 images from 19 students (12 expressions) | Gray processing, histogram equalization, & scale normalization | Extended LBP algorithm (ELBP) | Accuracy: 98.16% | Outperforms the original LBP algorithm | Focus only on general emotions |
| ScienceDirect | Hung et al., 2019 [39] | CNN | FER-2013, JAFFE and KDEF | AdaBoost | CNN | Accuracy: 84.59% | Achieved high recognition accuracy | Focus only on general emotions |
| IEEE Xplore | Lasri et al., 2019 [40] | CNN | FER-2013 | AdaBoost | CNN | Accuracy: 70% | Good in predicting happy and surprised emotions | Focus only on general emotions |
| Scopus | Tang et al., 2019 [41] | CNN | FER-2013 | Convert each $48 \times 48$ pixel grayscale face image into a string | CNN | Accuracy: 70.10% | (i) High detection accuracy and robustness<br>(ii) Real-time evaluation students' classroom performance | Focus only on general emotions |
| IEEE Xplore | Shi et al., 2019 [42] | CNN-SVM | 82 students that learn different online courses | Face detection using Viola-Jones, image rotation, image scaling normalization | CNN | Accuracy: 93.80% | High predictive performance | Only considers two levels of confusion |

Table 2: Continued.

| Retrieved from | Authors, year | Emotion classifier | Dataset | Preprocessing | Feature extraction | Results | Strength | Limitation |
|---|---|---|---|---|---|---|---|---|
| IEEE Xplore | Healy et al., 2019 [43] | SVM | CK+ and MUG | Dlib library, which contains CNNs trained in face detection. | Dlib is used to extract the 68 landmark points from the detected face | Accuracy: 88.76% | Able to provide quick and reliable classification | (i) Focus only on general emotions<br>(ii) Computationally intensive |
| Google scholar | Liang, 2019 [44] | SVM | Yale face of Kyushu University in Japan, BioID Face Database, JAFFE | Normalization of feature points | Active shape model (ASM) | Accuracy: 79.65% | Good recognition rate for some facial expressions | Focus only on general emotions |
| IEEE Xplore | Dash et al., 2019 [45] | CNN | DAiSEE | Viola-Jones algorithm | CNN | Accuracy:<br>(i)Engaged: 0.901 ± 0.09<br>(ii) Not engaged: 0.921 ± 0.04 | Achieved high detection accuracy | Focus only on engagement detection |
| Google Scholar | Bian et al., 2019 [46] | CNN | OL-SFED | VGG16 | CNN | Accuracy: 91.60% | Achieved high detection accuracy | Limited sample number |
| IEEE Xplore | Huang et al., 2019 [47] | BERN (combination of temporal convolution, bidirectional LSTM, and attention mechanism) | DAiSEE | N/A | Tracking the changes in face position and pixels through deep learning | Top 1 accuracy: 60% for four classification model | Achieved state-of-art performance compared with the benchmark (57.9%) | Requires a large amount of training data and a long training time |
| Springer | T. S and Guddeti, 2020 [48] | Hybrid CNN | Self-created dataset (engaged, boredom, and neutral) | Delete the blurred and repeated frames | N/A | Accuracy:<br>(i) Posed: 86%<br>(ii) Spontaneous: 70% | Outperforms the existing state-of-the-art methods | Long training time |
| Google Scholar | Mohamad Nezami et al., 2020 [49] | CNN | 4627 engaged and disengaged sample | CNN-based face detection algorithm | N/A | Precision: 60.42% | Outperforms their previous engagement recognition method | Focus only on engagement detection |
| Scopus | Alrayassi and Shilbayeh, 2020 [50] | CNN | 15 random students seeking admission in ADSM | AdaBoost | PCA algorithm | Accuracy: 56.60% | Able to detect happiness and no emotion | (i) Focus only on general emotions<br>(ii) Difficult to detect sadness and surprised emotion |
| Google Scholar | Tang et al., 2020 [51] | CNN | JAFFE and CK+ | CNN | CNN | Accuracy:<br>(i) JAFFE: 92.68%<br>(ii) CK+: 99.10% | Achieved high recognition accuracy | Focus only on general emotions |
| Google Scholar | Leong, 2020 [52] | LSTM | DAiSEE | MTCNN library—face detection & cropping | FaceNet model | Accuracy compared to the EmotioNet model:<br>(i) Boredom: +16.26%<br>(ii) Frustration -2.42% | (i) Improved accuracy for boredom emotion | (i) Decreased accuracy for frustration emotion<br>(ii) Involved negative emotions only |
| Springer | Zatarain Cabada et al., 2020 [53] | CNN | Database Insight (dbI) | CNN | Local binary patterns (LBP), geometric-based feature extraction, and convolutional filters (CF) | Accuracy: 82% | Demonstrate an 8% improvement in accuracy over a previous work that used a trial and error method | Imbalanced dataset |
| Springer | Zhu and Chen, 2020 [54] | Hybrid deep neural network (hybrid DNN) | JAFFE, CK+, and FER-2013 | Increase the original data by eight times, including the original image, flipped image, and rotated images of the six angles | Face++ detect API | Accuracy: 83.90% | Show high recognition accuracy | Requires a large amount of training data |
| Scopus | Wang et al., 2020 [7] | CNN | CK+, DISFA, DISFA+ | IntraFace | CNN | N/A | Performs robustly in various environments | Focus only on general emotions |
| IEEE Xplore | Dubbaka and Gopalan, 2020 [55] | CNN | DISFA+ | (i) Grayscale and reduced into $160 \times 224$ pixels input dimensions<br>(ii) Split into the upper, lower, and whole face | Using Dlib package | Accuracy: 95% | Achieved a high detection accuracy | Obstructions on faces limited the model's performance |

Table 2: Continued.

| Retrieved from | Authors, year | Emotion classifier | Dataset | Preprocessing | Feature extraction | Results | Strength | Limitation |
|---|---|---|---|---|---|---|---|---|
| Springer | Pise et al., 2020 [56] | Temporal relation network (TRN) | 30 samples of different individual's frontal images with four emotion types | Scale, align, and normalize the samples | Base (SqueezeNet) CNN | Accuracy: 91.30% | Achieved a high detection accuracy | (i) Prone to underfitting issues due to a small dataset<br>(ii) Focus only on general emotions |
| Google Scholar | Sabri, 2020 [57] | Support vector regression (SVR) | JAFFE | Viola-Jones algorithm | Gray-level cooccurrence matrix (GLCM) | Accuracy: 99.16% | (i) Less susceptible to overfitting<br>(ii) Achieved a high detection accuracy | (i) Focus only on general emotions<br>(ii) Computationally expensive |
| Scopus | Kumar et al., 2020 [58] | SVM | JAFFE, CK+, and FER-2013 | Kanade-Lucas-Tomasi algorithm | Gabor filter | Accuracy: 62% | Depression detection is included | Focus only on general emotions |
| IEEE Xplore | Murugappan et al., 2020 [59] | Extreme learning machine (ELM) and probabilistic neural network (PNN) | 55 undergraduate university students | Viola-Jones algorithm | Use a mathematical model to place ten virtual markers on the subject's face in defined locations | Accuracy:<br>(i) ELM: 88%<br>(ii) PNN: 92% | Achieved a high detection accuracy | Only simple distance measure is used for emotion classification |
| IEEE Xplore | Murugappan et al., 2020 [60] | K-nearest neighbors (KNN) & decision tree | 55 subjects with six types of emotions | Viola-Jones algorithm | Use a mathematical model to place ten virtual markers on the subject's face in defined locations | Accuracy: 98.03% | (i) Less computational complexity analysis<br>(ii) Achieved a high detection accuracy | Focus only on general emotions |
| Google Scholar | Rao and Rao, 2020 [61] | CNN | DAiSEE, JAFFE, and CK+ | (i) Each video is cut into frames<br>(ii) Resize to an image of size 48 × 48 pixels<br>(iii) Apply limited histogram equalization | CNN+pose estimator | Accuracy:<br>(i) DAiSEE: 53.4%<br>(ii) JAFFE: 71.4%<br>(iii) CK+: 99.95% | Achieved high detection accuracy when using the CK+ dataset | Show a low recognition rate for frustration when using the DAiSEE dataset |
| Google Scholar | Hingu, 2020 [62] | CNN | FER-2013 | Haar cascades | CNN | Accuracy:<br>(i) Training set: 65%<br>(ii) Validation set: 63% | The feature extraction method outperforms the existing traditional approach | (i) Focus only on general emotions<br>(ii) Low detection accuracy |
| IEEE Xplore | Zakka and Vadapalli, 2020 [63] | CNN | FER-2013 | Haar cascades | CNN | Accuracy: 64.43% | Able to detect emotions in real time | (i) Focus only on general emotions<br>(ii) Low recognition accuracy |
| Springer | Liao et al., 2021 [64] | Deep facial spatiotemporal network (DFSTN) | DAiSEE | MTCNN | SE-ResNet-50 (SENet) | Accuracy: 58.84% | (i) The prediction effect is enhanced even in challenging circumstances<br>(ii) Able to fuse facial spatiotemporal information | (i) Low detection accuracy<br>(ii) Data deficiencies and data imbalances |
| IEEE Xplore | Siam et al., 2021 [65] | CNN | FER-2013 | (i) Resize images into 224 × 224<br>(ii) Image augmentation<br>(iii) Normalization | CNN | Accuracy: 69% | Able to generate reviews from an image with multiple faces | (i) Focus only on general emotions<br>(ii) Low detection accuracy |
| Google Scholar | Li et al., 2021 [66] | CNN | FER-2013 | Haar cascades | CNN | Accuracy: 72.4% | Less complex model | Focus only on general emotions |
| IEEE Xplore | Mohan et al., 2021 [67] | Deep CNN (DCNN) | FER-2013, JAFFE, CK+, KDEF, and RAF | (i) Rotation by $+5^{\circ}$<br>(ii) Rotation by $-5^{\circ}$<br>(iii) Horizontal flip<br>(iv) Gaussian noise | DCNN | Accuracy:<br>(i) FER-2013: 78%<br>(ii) JAFFE: 98%<br>(iii) CK+: 98%<br>(iv) KDEF: 96%<br>(v) RAF: 83% | Outperforms twenty-five baseline methods by considering the average time | The performance is generally not as good as that in FER under a lab-controlled environment |

Table 2: Continued.

| Retrieved from | Authors, year | Emotion classifier | Dataset | Preprocessing | Feature extraction | Results | Strength | Limitation |
|---|---|---|---|---|---|---|---|---|
| Springer | Mohan et al., 2021 [68] | CNN | FER-2013, JAFFE, CK+, KDEF, and RAF | Image resizing using bilinear interpolation | CNN | Accuracy: (i) FER-2013: 78.9% (ii) JAFFE: 96.7% (iii) CK+: 97.8% (iv) KDEF: 82.5% (v) RAF: 81.68% | More preeminent in terms of accuracy and execution time compared to 21 state-of-the-art methods | Focus only on general emotions |
| Springer | Mohan and Seal, 2021 [69] | SVM, RF, KNN, MLP, Adaboost | Real Life dataset (RL) and Bag-of-Lies | (i) The facial regions of the selected frames are cropped (ii) Images are reshaped using bilinear interpolation (iii) LDP face images are concatenated and resized using bilinear interpolation | N/A | Accuracy: (i) Bag-of-Lies (video data +audio data +EEG signals data): 70% (ii) RL dataset (video data +audio data): 76.07% | Combining modalities are consistent with deception detection | Small datasets used |
| IEEE Xplore | Mohan et al., 2022 [70] | Deep CNN (DCNN) | RL trail, Bag-of-Lies, MU3D | (i) The facial regions of the selected frames are cropped (ii) Images are reshaped using bilinear interpolation (iii) LDP face images are concatenated and resized using bilinear interpolation | DCNN | Accuracy: (i) RL trail: 97% (ii) Bag-of-Lies: 96% (iii) MU3D: 98% | Achieved a high detection accuracy | Data scarcity |
| Springer | Shen et al., 2022 [71] | Squeeze and excitation-deep adaptation networks (SE-DAN) | JAFFE, CK+, and RAF-DB | Random rotation and horizontal flip on RAF-DB | SE-CNN | Accuracy: 56% | (i) The accuracy is higher than Alexnet, VGG-16, SE-CNN, and DAN (ii) Can be used for transfer learning and domain adaptation | Focus only on general emotions |
| Springer | Gupta et al., 2022 [72] | DCNN (ResNet-50, VGG19, Inception-V3) | FER-2013, CK+, RAF-DB, and own dataset (consists of 1800 coloured images with emotions such as angry, sad, happy, neutral, surprised, and fear) | Automatic frame selection | MediaPipe face mesh | Accuracy: (i) ResNet-50: 92.32% (ii) VGG19: 90.14% (iii) Inception-V3: 89.11% | Outperforms all other models for FER in real-time learning scenarios | Focus only on engagement detection |
| IEEE Xplore | Savchenko et al., 2022 [73] | CNN | AffectNet | Rotate cropped facial images to align them based on eyes position | CNN | Accuracy: 70.23% | Much faster and can be implemented for real-time processing | Less accurate when compared to the best-known multimodal ensembles on the AFEW and VGAF datasets |
| Google Scholar | Hou et al., 2022 [74] | CNN (VGG16) | FER-2013 and CK+ | MTCNN | VGG16 | Accuracy: (i) FER-2013: 67.4% (ii) CK+: 99.18% | The accuracy of VGG16 +ECANet is 2.76% higher than VGG16 itself | Slow algorithm's running speed |
| Google Scholar | Yuan, 2022 [75] | MTCNN | RAF-DB, masked dataset, and classroom dataset | Histogram equalization | MTCNN | Accuracy: 93.53% | (i) Able to detect multiple faces in a single image (ii) The pressure of data storage is alleviated, and the collection workload is reduced | Small sample size of the dataset |
| Google Scholar | Wu, 2022 [76] | CNN | The self-collected dataset consists of 1073 images with expressions such as boredom, surprise, happy, confusion, and neutral | (i) Convert pictures into grayscale (ii) Resize into 160 × 160 pixels | CNN | Accuracy: 80% | Perform well in recognizing emotions such as happy, surprised, and neutral | Dataset insufficiency |

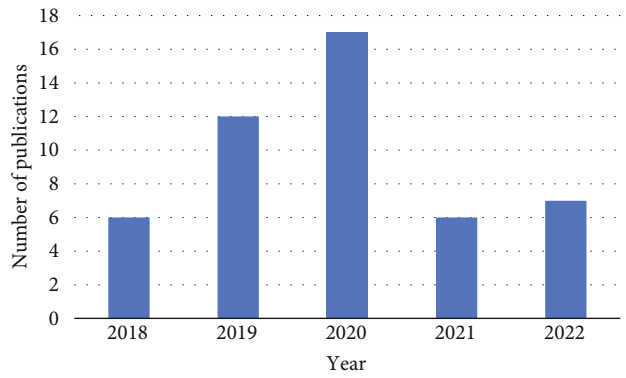


Figure 4: The year-wise distribution for 48 primary studies.

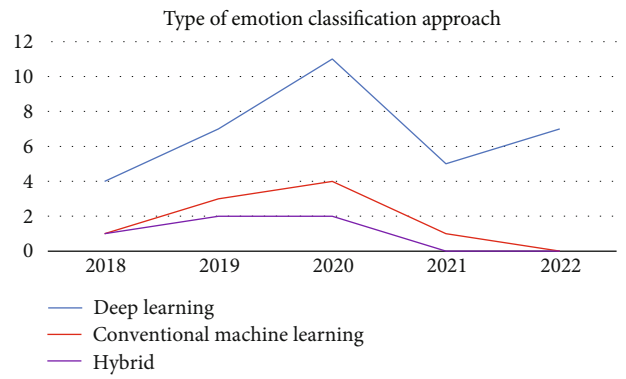


Figure 5: The type of approach used to classify emotion.

in the research publications that were evaluated over the years. It was discovered that most papers relevant to the topic of this review paper were published in the year 2020. In recent years, most studies (70.83%) applied deep learning algorithms to classify emotions, while 18.75% used machine learning algorithms and 10.42% applied hybrid algorithms. Figure 5 illustrates the classification approach used in the previous studies from 2018 to 2022.

The type of approach used in the retrieved publications was investigated and extracted into Table 2 to address the first research question (RQ1). Besides that, the dataset used in the reviewed publications was also summarized in Table 2 to answer RQ2. FER-2013 is the most used dataset in the FER system. Besides that, popular datasets such as CK+ and JAFFE were also used in previous studies. The dataset focusing on academic emotions, such as DAiSEE, was also used in seven publications out of 48 primary studies.

Furthermore, it was found that CNN is the most frequently used deep learning algorithm as an emotion classifier in the FER systems. On the other hand, SVM is commonly used in conventional machine learning-based FER systems. The percentage of real time and non-real-time FER systems are summarized in Figure 6 to answer RQ4. The percentage of non-real-time FER systems was found to be higher than the real-time FER systems. Other than that, DAiSEE and OL-SFED are examples of datasets that consist of academic emotions. The overall analysis of the FER in online learning was performed by comparing the crucial components in each relevant research study, such as dataset, emotion classifier, preprocessing method, feature extraction approach, results of the research experiment, and strengths and limitations.

According to the analysis of research papers collected for this SLR, 34 papers used deep learning algorithms as the emotion classifier in their research experiment. Out of 34 studies, a study conducted by Rao et al. in 2020 achieved the highest accuracy of 99.95% using CNN as the emotion classifier and CK+ as the dataset [61].

Furthermore, nine papers applied conventional machine learning algorithms for emotion classification. The study conducted by Sabri et al. in 2020 achieved the highest accuracy of 99.16% using SVR as the emotion classifier and JAFFE as their dataset [57]. Out of five papers that used hybrid algorithms in facial emotion classification, a study by Shi et al. in 2019 achieved the highest accuracy of 93.80% [42]. They used a combination of CNN and SVM as their emotion classifier, and the dataset, which consists of 82 students that learn different online courses.

Table 2 includes the strengths and limitations of the papers, which are not confined solely to the strengths and limitations of the emotion classifier but also encompass the datasets used in these studies. Hence, a summary of the advantages and disadvantages of each emotion classifier algorithm will be presented in Table 3.

Ultimately, a quantitative comparison may be preferable when the objective is to select the model with the best performance on classification tasks or to identify the model with the highest accuracy. However, it is essential to note that quantitative measures cannot necessarily capture all the relevant aspects of a model. Hence, qualitative and quantitative factors could be significant in picking the ideal FER model. A summary of the FER classification methods is presented in Table 4.

## 5. Discussion

In this section, the commonly used datasets and academic datasets that were used in the selected reviewed publications were discussed. Besides that, the conventional learning emotion algorithms and deep learning algorithms applied were also presented in detail. Finally, this section is concluded with a critical review where the publications that used academic emotion datasets and deep learning algorithms were reviewed.

*5.1. Commonly Used FER Datasets.* There are multiple online datasets available for the FER field, including FER-2013, JAFFE, CK+, KDEF, DISFA, and DISFA+. Most datasets available are constructed based on 2D video sequences or static images. Nonetheless, the 3D image can be found in certain datasets. The six basic emotions, including neutral, happiness, disgust, anger, fear, surprise, and sadness, are labeled in most datasets. Some datasets are built in controlled environments, while others are created in wild environments. This section presents several well-known and

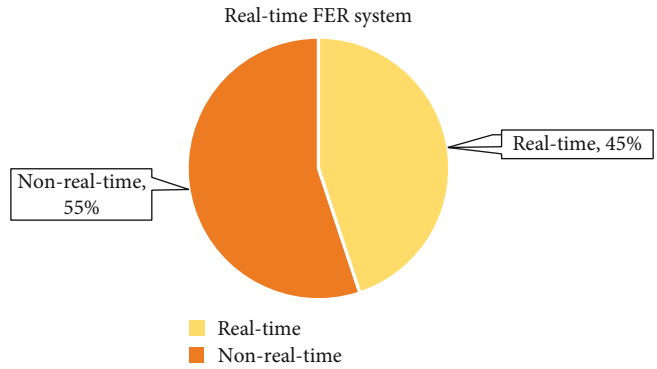


Figure 6: The percentage of real-time FER systems in the retrieved publications.

commonly used datasets in the reviewed works. The commonly used FER datasets are summarized in Table 5.

*5.1.1. FER-2013.* FER-2013 is a dataset generated using the Google Search API, which consists of six basic emotions and neutral emotions by matching 184 keywords related to emotion [91]. Detailed information on the race or ethnicity of the individuals is not provided in the dataset. It consists of roughly 30,000 grayscale and 48x48 scaled facial images with various facial expressions, and the main labels of it can be divided into seven types. FER-2013 is the largest publicly accessible dataset for facial emotion in the wild. Nevertheless, it is challenging for facial landmark detectors to extract landmarks due to the image resolution and quality.

*5.1.2. Japanese Female Facial Expression (JAFFE).* JAFFE is a dataset of 213 images of facial expressions from ten Japanese female individuals [92]. Every female had to make seven facial expressions, and 60 annotators annotated the images with average semantic scores for each facial expression. Each facial image has an image resolution of $256 \times 256$ pixels.

*5.1.3. Extended Cohn-Kanade (CK+).* The CK+ dataset comprises 593 video sequences, covering 123 different participants with different genders and heritage aged between 18 to 50 years old [93]. There were 81% Euro-Americans, 13% Afro-Americans, and 6% from other groups. 69% of them were female, while 31% were male. Each video depicts a transformation in facial emotion from neutral to specific peak emotion, captured at 30 FPS with a resolution of either $640 \times 480$ or $640 \times 490$ pixels. There are 327 videos labeled with seven classes of facial expressions, including happiness, surprise, disgust, sadness, anger, fear, and contempt.

*5.1.4. Karolinska Directed Emotional Faces (KDEF).* This dataset contains 4900 photos of facial emotions from 70 individuals, captured from five different angles and labeled with six basic facial expressions plus neutral [94]. The subjects are 35 males and 35 females aged between 20 and 30 years old. This dataset was created in Sweden and is designed to represent Sweden's population. The photographs of individuals from various backgrounds and ethnicities were included, but the exact demographics of the individuals are not explicitly stated on the official dataset website. During the photography session, subjects without beards, eyeglasses, earrings or moustaches, as well as no noticeable make-up, are preferred. Each image has a resolution of $562 \times 762$ pixels.

*5.1.5. Denver Intensity of Spontaneous Facial Action Database (DISFA).* This dataset contains spontaneous facial emotions that can be used to automatically detect action units and intensities described by FACS [95]. This dataset includes videos of 12 females and 15 males of various ethnic groups. Twenty-one were Euro-American, three were Asian, two were Hispanic, and one was African-American. Sixty-six facial landmark points are included in each picture. The pictures in the DISFA dataset were captured at a high resolution ($1024 \times 768$) using the PtGrey stereo imaging system.

*5.1.6. Extended Denver Intensity of Spontaneous Facial Action Database (DISFA+).* The DISFA+ database is an extension of the DISFA database and comprises a massive amount of data on posed and spontaneous facial emotions [96]. The participants in the DISFA+ dataset are from the same population as those in the original DISFA dataset. Besides that, it also includes metadata as well as manually labeled frame-based annotations of 5-level intensity for 12 FACS facial actions.

*5.2. Academic Emotion Datasets.* Besides FER datasets focusing on seven facial expressions (6 basic expressions and one neutral expression), only a few publicly accessible academic emotion datasets were recorded in an online learning environment. This section will discuss some academic emotion datasets.

*5.2.1. Dataset for Affective States in E-Environments (DAISEE).* This dataset comprises 9068 video clips from 112 Indian students to detect the individual affective states of frustration, engagement, confusion, and boredom [97]. The videos in the dataset are captured in dormitories, laboratories, and crowded classrooms where the students focus on their educational tasks on the computer screen. The annotations will be further measured according to the intensity, which is from 0 to 3.

*5.2.2. Online Learning Spontaneous Facial Expression Database (OL-SFED).* OL-SFED is a dataset containing 30184 images and 1274 video clips by 82 Chinese students

Table 3: Summary of the advantages and disadvantages of the emotion classifier algorithms.

| FER classification method | Algorithm | Advantage | Disadvantage |
|---|---|---|---|
| Conventional machine learning algorithm | Viola-Jones algorithm | Fast face detection algorithm [34] | Has lower detection accuracy compared to more complex algorithms [34] |
| | Support vector machine (SVM) | High accuracy in classification tasks [77] | Computationally intensive, especially when dealing with large datasets or complex models [77] |
| | Support vector regression (SVR) | Less susceptible to overfitting [57] | Computationally expensive, particularly for large datasets [78] |
| | Extreme learning machine (ELM) | Perform faster in classification [59] | Takes more computational time and has a lower accuracy than PNN [59] |
| | Probabilistic neural network (PNN) | Takes less computational time than ELM and is more efficient in classification [59] | Possible overfitting of the data [59] |
| | Decision tree | Able to handle missing data by not incorporating the missing feature during the decision-making process [79] | More computation time compared to KNN [60] |
| | K-nearest neighbors (KNN) | Achieved higher accuracy when compared to the decision tree algorithm and lesser computation time [60] | Slower performance compared to decision tree [60] |
| | Multilayer perceptron (MLP) | Capable of adaptive learning and optimal processing [80] | Lower classification accuracy compared to the random forest algorithm [69] |
| | Adaptive boosting (AdaBoost) | Enhance the performance of classification out of weak learners [81] | Susceptible to noisy data [82] |
| | Random forest (RF) | Robust to noisy data or outliers [83] | Prone to overfitting [84] |
| Deep learning algorithm | Convolutional neural networks (CNN) | Effective at handling complex image and video data [61] | Requires a large amount of training data and significant data augmentation to avoid overfitting [85] |
| | Neural network | Able to achieve high classification accuracy [33] | Long processing time [33] |
| | Deep belief network (DBN) | Robustness in classification [35] | Requires large amounts of training data [35] |
| | Long short-term memory (LSTM) | Addressed the issue of vanishing gradients [52] | Slow computational speed for complex architectures [52] |
| | Temporal relation network (TRN) | Able to achieve state-of-the-art performance on FER benchmarks [56] | Prone to overfitting/underfitting on small datasets [56] |
| | Deep facial spatiotemporal network (DFSTN) | Able to fuse facial spatiotemporal information [64] | Require larger amounts of training data to learn effective feature representations and to avoid overfitting [64] |
| | Deep CNN (DCNN) | Effective at learning complex features from raw image data [86] | Difficult to interpret, as it might be challenging to understand the underlying mechanisms behind the model's decision-making process [87] |

Table 3: Continued.

| FER classification method | Algorithm | Advantage | Disadvantage |
|---|---|---|---|
| | Squeeze and excitation-deep adaptation networks (SE-DAN) | Can be used for transfer learning and domain adaptation [71] | Require a significant amount of computational resources and time to train [71] |
| | Multitask cascaded convolutional neural network (MTCNN) | High accuracy in detection and classification tasks; able to detect multiple faces in a single image [75] | Require a large amount of training data to achieve high accuracy [75] |
| Hybrid algorithm | Support vector machine +convolutional neural network (SVM +CNN) | Enhance the performance of classification compared to using just one of these algorithms alone [30] | Hyperparameter tuning of this combination of two algorithms can be challenging and time-consuming [88] |
| | BERN (combination of temporal convolution, bidirectional LSTM, and attention mechanism) | Achieved state-of-art performance [47] | Requires a large amount of training data and a long training time [47] |
| | Hybrid convolutional neural network (hybrid CNN) | More robust to variations in input data [89] | Long training time [48] |
| | Hybrid deep neural network (hybrid DNN) | Able to handle a wide range of data types and classification tasks [90] | Requires a large amount of training data [54] |

in an online learning environment [46]. The dataset comprises spontaneous facial expressions in response to 5 typical academic emotions, including enjoyment, fatigue, neutral, distraction, and confusion, with samples thoroughly annotated by participants and external coders.

*5.3. Preprocessing Methods.* According to the selected 48 retrieved publications, most of the studies applied the Viola-Jones algorithm, also known as the Haar cascades classifier, during the preprocessing phase.

The viola-Jones algorithm is the most commonly implemented face detection algorithm that searches through an image using a window to seek features of a human face [98]. A face is assumed to be included inside a certain window of an image if and only if these features are identified and assigned a value that is unique to faces. This method consists of four steps, as presented in Figure 7.

The first step that has to be done in this method is to read the image of the person's face that is facing the camera [99]. Next, the Haar-like feature will interpret the camera-captured facial image by processing the image into boxes to signify dark and bright areas of the facial image. Some features of the human face are universal such as the fact that the area around the eyes is always darker than its surrounding pixels, and the area around the nose is always lighter.

The following step is to compute all of the pixels that are contained within that specific feature. The calculation of an integral image can be carried out in a time-efficient and effective manner for each and every point in the images. After obtaining the integral image of all points, the pixel intensity of every subwindow in the image may be calculated with a maximum of four memory references. Adaptive Boosting, or AdaBoost, is one of the most popular boosting approaches that merges a number of underperforming classifiers to create a strong classifier. It chooses several weak classifiers to combine into a single model, giving each classifier weight to produce a robust classifier.

The final step in the Viola-Jones method is the cascade classifier. The Haar cascades are a kind of classifier that can detect an object in a previously trained image or video [100]. Each subwindow will have a specific feature assigned to it in order to determine its classification in the initial stage of the classification. The output of the feature is considered to be rejected if it does not satisfy the desired requirements. After that, the algorithm will go to the subsequent subwindow, where it will do another calculation to determine the value of the feature. The algorithm will proceed with the subsequent phase once the acquired result meets the prerequisite criteria. Finally, it is regarded to contain a face if a subwindow is able to pass through all the phases of the classification process.

Based on the retrieved publications, Ayvaz et al. [101], Candra Kirana et al. [34], Dewan et al. [35], El Hammoumi et al. [31, 32], Shi et al. [42], Dash et al. [45], Sabri [57], Murugappan et al. [59], and Murugappan et al. [60] have applied the Viola-Jones algorithm in the preprocessing step. Moreover, Ma et al. [32], Yang et al. [33], Hingu [62], and Zakka and Vadapalli [63] have used the Haar cascades method, while Hung et al. [39], Lasri et al. [40], and Alrayassi and Shilbayeh [50] have implemented the AdaBoost algorithm in the preprocessing phase.

Apart from that, histogram equalization is also applied in preprocessing stage by G. Li and Wang [30], Mao et al. [38], and Rao et al. [61]. Histogram equalization is a computer image processing method applied to enhance image contrast. This method is applied because of the specialized nature of learning activities, where there is more light in the scene and less variation in the learner's head angle. Besides the commonly used preprocessing method, Kumar et al. [58] applied the Kanade-Lucas-Tomasi algorithm for

Table 4: Summary of FER classification methods.

| FER classification method | Algorithm | Author, year |
|---|---|---|
| Conventional machine learning algorithm | Viola-Jones algorithm | Candra Kirana et al., 2018 [34] |
| | Support vector machine (SVM) | Mao et al., 2019 [38] |
| | | Healy et al., 2019 [43] |
| | | Liang, 2019 [44] |
| | | Kumar et al., 2020 [58] |
| | | Mohan and Sael, 2021 [69] |
| | Support vector regression (SVR) | Sabri, 2020 [57] |
| | Extreme learning machine (ELM) | Murugappan et al., 2020 [59] |
| | Probabilistic neural network (PNN) | |
| | Decision tree | Murugappan et al., 2020 [60] |
| | K-nearest neighbors (KNN) | Murugappan et al., 2020 [60] |
| | | Mohan et al., 2021 [69] |
| | Multilayer perceptron (MLP) | Mohan and Seal, 2021 [69] |
| | Adaptive boosting (AdaBoost) | |
| | Random forest (RF) | |
| Deep learning algorithm | Convolutional neural networks (CNN) | El Hammoumi et al., 2018 [31] |
| | | Ma et al., 2018 [32] |
| | | Sharma and Mansotra, 2019 [37] |
| | | Sharma and Mansotra, 2019 [36] |
| | | Hung et al., 2019 [39] |
| | | Lasri et al., 2019 [40] |
| | | Tang et al., 2019 [41] |
| | | Dash et al., 2019 [45] |
| | | Bian et al., 2019 [46] |
| | | Mohamad Nezami et al., 2020 [49] |
| | | Alrayassi and Shilbayeh, 2020 [50] |
| | | Tang et al., 2020 [51] |
| | | Zatarain Cabada et al., 2020 [53] |
| | | Wang et al., 2020 [7] |
| | | Dubbaka and Gopalan, 2020 [55] |
| | | Rao & Rao, 2020 [61] |
| | | Hingu, 2020 [62] |
| | | Zakka and Vadapalli, 2020 [63] |
| | | Siam et al., 2021 [65] |
| | | Li et al., 2021 [66] |
| | | Mohan et al., 2021 [68] |
| | | Savchenko et al., 2022 [73] |
| | | Hou et al., 2022 [74] |
| | | Wu, 2022 [76] |
| | Neural network | Yang et al., 2018 [33] |
| | Deep belief network (DBN) | Dewan et al., 2018 [35] |
| | Long short-term memory (LSTM) | Leong, 2020 [52] |
| | Temporal relation network (TRN) | Pise et al., 2020 [56] |
| | Deep facial spatiotemporal network (DFSTN) | Liao et al., 2021 [64] |
| | Deep CNN (DCNN) | Mohan et al., 2021 [67] |
| | | Mohan et al., 2022 [70] |
| | | Gupta et al., 2022 [72] |

Table 4: Continued.

| FER classification method | Algorithm | Author, year |
|---|---|---|
| | Squeeze and excitation-deep adaptation networks (SE-DAN) | Shen et al., 2022 [71] |
| | Multitask cascaded convolutional neural network (MTCNN) | Yuan, 2022 [75] |
| Hybrid algorithm | Support vector machine +convolutional neural network (SVM +CNN) | Li and Wang, 2018 [30] |
| | | Shi et al., 2019 [42] |
| | BERN (combination of temporal convolution, bidirectional LSTM, and attention mechanism) | Huang et al., 2019 [47] |
| | Hybrid convolutional neural network (hybrid CNN) | T. S and Guddeti, 2020 [48] |
| | Hybrid deep neural network (hybrid DNN) | Zhu and Chen, 2020 [54] |

Table 5: Summary of commonly used FER datasets.

| Dataset | Type of emotions | Data configuration |
|---|---|---|
| FER-2013 | 6 basic facial emotions+1 neutral emotion | (i) Roughly 30,000 grayscale images<br>(ii) 48×48 scaled facial images with various facial expressions |
| JAFFE | 6 basic facial emotions +1 neutral emotion | (i) 213 images of various facial expressions<br>(ii) Ten different Japanese females<br>(iii) Image resolution of $256 \times 256$ |
| CK+ | Happiness, surprise, disgust, sadness, anger, fear, and contempt | (i) 593 video sequences<br>(ii) 123 different participants with different genders and heritage aged between 18 to 50 years old<br>(iii) Captured at 30 FPS with a resolution of either $640 \times 480$ or $640 \times 490$ |
| KDEF | 6 basic facial emotions+1 neutral emotion | (i) 4900 photos of human facial emotions from 70 individuals<br>(ii) Captured from five different angles<br>(iii) 35 males and 35 females aged between 20 to 30 years old<br>(iv) Image resolution of $562 \times 762$<br>(v) Subjects without beards, eyeglasses, earrings or moustaches, and no noticeable make-up |
| DISFA | Intensity of 12 AUs coded | (i) Stereo videos of 12 females and 15 males of various ethnic groups<br>(ii) Image resolution of $1024 \times 768$<br>(iii) 66 facial landmark points |
| DISFA+ | 5-level intensity of twelve FACS | (i) Extension of the DISFA database<br>(ii) Manually labeled frame-based annotations of 5-level intensity for 12 FACS facial actions |

face detection in the preprocessing phase. This algorithm has achieved a high true-positive rate under various exposure settings, and it can effectively detect facial parts like the eyes, nose, and mouth. Next, Wang et al. [7] introduced Infra-Face, a preprocessing tool that integrates algorithms for facial attribute detection, head pose estimation, facial feature tracking and more. As a result, the face's essential features, such as mouth, nose tip, eyes and eyebrows, are easily detected, and the emotions can be recognized by rectangular outlines accordingly.

In the study by Krithika and Priya [5], colour conversion is used in the preprocessing phase. Firstly, the RGB image is transformed into the $L * a * b$ colour space. Then, the a and b colour spaces are chosen and transformed into binary images. Finally, on the image, the AND operation is carried out in order to produce a matrix with the values 0 and 1, where 0 represents the absence of a face while 1 represents the existence of a face. Meanwhile, Pise et al. [56] performed face detection and preprocessing, such as scaling, aligning, and normalizing the samples. According to their research findings, alignment and normalization of the input samples can assist the deep neural network in learning relevant facial features. Besides that, there is another preprocessing technique, like the conversion of a grayscale image into a string,

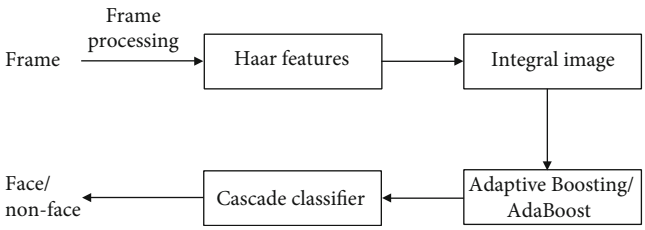


Figure 7: Schematic of Viola-Jones method.

performed by Tang et al. [41]. The grayscale image of the face, which is 48 by 48 pixels in size, is converted into a string for each individual image. This technique can help reduce the dimension and model complexity.

In conclusion, preprocessing enhances the performance of facial emotion recognition since it helps in reducing the noise that is present in the images. Therefore, it is an essential stage in the image processing process in the computer vision field.

*5.4. Feature Extraction Methods.* Based on the selected 48 retrieved publications, most of the publications applied a convolutional neural network (CNN) in the feature extraction phase.

Feature extraction part and classification part are the two fundamental components that make up a CNN. Multiple layers of convolutional and pooling layers make up the feature extraction network. In image processing, convolution is an effective feature extraction method adept at lowering the dimensionality of data and yielding a less redundant data set, also known as a feature map [102]. In addition, each kernel acts as an identifier for a feature and may thus filter out locations in the original picture where the feature is present. Eventually, it will generate a map, and the altitude on that map displays how these features are distributed.

A convolutional layer condenses the input data by extracting interest features within it and then constructing feature maps in response to various feature detectors [102]. By condensing the input and extracting features of interest, a convolutional layer can then provide feature maps that various feature detectors can use. The neurons in the first convolutional layer are responsible for filtering out basic features such as edges. High-order feature detection is achieved as a result of the neurons learning to collect information in order to obtain a more comprehensive view of the image in the subsequent convolutional layers. The most common pooling techniques utilized in CNNs are known as max-pooling and average-pooling. The purpose of pooling is to reduce the dimensionality of data in order to minimize overfitting by concentrating local data using a pooling window. Additionally, appropriate pooling results in invariance with respect to translation, scale, and rotation, as minor dislocations or scalings no longer have an effect.

Based on the selected reviewed publications, it was found that El Hammoumi et al. [31], Ma et al. [32], Hung et al. [39], Lasri et al. [40], Tang et al. [41], Shi et al. [42], Dash et al. [45], Bian et al. [46], Tang et al. [51], Wang et al. [7], Pise et al. [56], Hingu [62], and Zakka and Vadapalli [63] have applied CNN for the feature extraction.

Besides CNN, local binary pattern (LBP) was also used in the studies conducted by Mao et al. [38] and Zatarain Cabada et al. [53]. It is feasible to define the shape of a digital image using LBP as well as the texture of a digital image [103]. This is accomplished by first segmenting the image into multiple small parts, from which the features are extracted. Furthermore, Dewan et al. [35] made use of a feature extraction technique known as Local Directional Pattern (LDP) to extract person-independent edge features for the various facial emotions. It is robust enough to generate consistent representation even when nonmonotonic illumination change and random noise are present.

Other than that, Ayvaz et al. [101] proposed facial landmark localization as the feature extraction method in their study. Facial landmarking refers to the process of detecting and locating specific points and features on a person's face. The algorithm generates 68 facial landmarks on the face that have been detected, each of which indicates the boundaries between different facial features.

Wibawanto and Kirana [6] applied median fisher's face in their research, a technique that combines the linear feature extraction and reduction approaches of princpal component analysis (PCA) and linear discriminant analysis (LDA). This proposed approach employs LDA in PCA space after detecting the face to obtain the fisher's face. Following that, the fisher face is transformed into the fisher's median. Moreover, Haar cascade extraction was used in the previous studies conducted by El Hammoumi et al. [31, 32]. Based on the research findings, the Haar cascade extraction approach would be the ideal method for high-performance images and high-resolution faces. This is because the Gabor filter and wavelet transform only extract particular facial features, leaving out those when the face is not correctly aligned facing the camera.

Besides that, Liang [44] proposed an enhanced active shape model (ASM) approach, one of the most used face feature point localization algorithms for extracting face feature points. This approach is not overly complicated and can be comprehended with little effort. However, there is room for improvement in this technique's speed, accuracy, and overall effectiveness with regard to applications. According to the studies conducted by Alrayassi and Shilbayeh [50], Principal Component Analysis (PCA) was used for the facial feature extraction. Besides feature extraction, dimensionality reduction is also included by applying the PCA algorithm.

Next, Kumar [58] applied the Gabor filter to identify the key features of the faces. Utilizing a Gabor filter bank makes it possible to extract various features. Zhu and Chen [54] used Face++ Detect API to accurately extract 106 facial landmark points, including 20 landmark points of the mouth, 20 for two eyes, 15 for the nose, 18 for eyebrows, and 33 for facial contour. Furthermore, the gray-level cooccurrence matrix (GLCM) was used for the feature extraction in the research made by Sabri [57]. GLCM is used to train the grayscale images of the eyes and mouth regions for texture analysis.

To sum up, feature extraction is one of the crucial phases in FER because it determines the overall efficiency of the FER process. Furthermore, since this phase can minimize

the dimensionality of data by eliminating redundant data, it can facilitate faster training and inference, which in turn increases the accuracy of learned models.

*5.5. Conventional Learning Emotion Classifier.* In 2016, a student emotion recognition system for online learning was presented in the previous study [5]. It can capture the students' emotions that are dynamically shifting when listening to the lecture in the e-learning environment. The local binary patterns (LBP) and Viola-Jones methods were applied for the detection of students' faces and the classification of students' emotions. The authors claimed that the quality could be accomplished better by using eye and head movements based on the concentration level recognized.

A facial emotion recognition system was developed in 2017, which detects the students' emotional states and motivation in the video conference form of e-learning based on facial expressions [101]. Machine learning approaches such as SVM, KNN, CART, and random forest were applied in the classification, and SVM and KNN algorithms achieved the best accuracy rates.

Besides that, a facial recognition model that detects emotion in a virtual learning environment was proposed in 2018 [33]. The authors used the Haar cascades approach to detect the input image to extract mouth and eyes on the JAFFE dataset to detect emotions. The average accuracy of each emotion is 82.58% for fear, 84.32% for disgust, 91.22% for anger, 95.25% for happiness, 93.26% for surprise, and 78.54% for sadness. However, the authors stated there was uncertainty on how the image's illumination and pose would influence the final emotion recognition because they do not involve those two factors. This section will further discuss commonly used conventional learning emotion classifiers, including SVM, decision tree, and random forest.

*5.5.1. Support Vector Machine (SVM).* This algorithm is a common regression and classification prediction algorithm, which applies machine learning theory to boost prediction accuracy while avoiding data overfitting. Despite that, since it is crucial to choose the proper kernel function, they are costly, difficult to tune, and do not efficiently perform with large databases. It is also known as a system that applies linear function hypothesis space in a high-dimensional feature space and is trained using an optimization theory-based learning approach that includes a statistical learning bias [104].

*5.5.2. Decision Tree.* Similar to SVM, the decision tree is also typically applied in regression and classification tasks, mainly in classification. Data features are represented as nodes in a tree-like structure, and each branch symbolizes the decision rule while each leaf node indicates the result. The threshold value of neutral emotion is used in generating the decision tree that provides rules to classify the data. The tree, in the form of a top-down method, is constructed without backtracking that determines how to make decisions in order to detect various facial expressions [105].

*5.5.3. Random Forest.* This algorithm is a classification algorithm containing a large number of individual decision trees. Every decision tree produces a prediction, and a majority vote decides the final prediction, which ensures that the last prediction is the most predicted class. Overfitting is one of the major machine learning problems but can be prevented using a random forest classifier. The overfitting problem will not occur as long as sufficient trees are available in the forest [106]. In contrast, many trees in the random forest caused the algorithm to be slow and inefficient for predictions in real time.

*5.6. Deep Learning Emotion Classifier.* A deep learning approach reduces the dependency on image preprocessing and feature extraction in terms of FER [107]. It is robust in environments with different components, such as occlusion or illumination, which enables them to outperform the conventional approaches. Furthermore, the deep learning approach is capable of handling large datasets. In this section, several commonly applied deep-learning emotion classifiers will be further discussed.

*5.6.1. Convolutional Neural Network (CNN).* CNN has accomplished state-of-the-art in multiple fields such as FER, face recognition, and object recognition because the physics-based model's dependency or other preprocessing approaches can be highly reduced or completely removed by enabling "end-to-end" learning from input pictures [15]. The main advantage of CNN is that there is only a slight effect on the recognition effect on the geometric transformation, deformation, and illumination.

A CNN is made of convolution, pooling, and fully connected layers [108]. The convolution layer can extract feature that combines linear and nonlinear operations, such as activation function and convolution operation. In order to let the neural network learn efficiently with high accuracy, the linear combination of features is transformed into nonlinear by the activation function. Next, the dimensionality of each function map dimensionality is reduced by the pooling layer while keeping the important information. The images are then partitioned into overlapping or nonoverlapping regions, and each region is downsampled by a nonlinear function like max-pooling and average pooling. After that, the output feature maps of the final convolution or pooling layer are often flattened and connected to at least one fully connected layer, in which every input is associated with every output by a learnable weight.

CNN has the benefit of automatically detecting the essential features without human intervention, as opposed to its predecessors. In addition, CNN is computationally effective since special convolution and pooling operations are used, and the parameters are shared. This allows the functioning of CNN models on any device, which makes them universally attractive. Moreover, CNN is great at classifying two similar emotions because it processes more granular elements within an image.

Many recent studies have used deep learning algorithms such as CNN to infer emotions. One of them is a facial emotion recognition model that combines CNN with specific image preprocessing steps [109]. CNN was used to detect seven basic emotions. In addition, 96.76% accuracy on the

CK+ dataset was obtained using preprocessing techniques such as the generation of synthetic samples and normalization of intensity and spatial. Nevertheless, the authors stated that their study is limited to the subject's frontal face of the controlled environment's input images.

Furthermore, Ma et al. presented an emotion recognition model using CNN [32]. The proposed model captures students' images through a web camera, analyzes students' real-time learning emotions, and gives the lecturer feedback. The authors found that lecturers who used the proposed model instinctively picked up on their students' emotional states, in contrast to lecturers who did not use it. Nevertheless, some lecturers refuse to use the emotional analysis model as the sentiment scores updating in real-time obstruct their teaching concept.

In 2019, a deep learning-based facial emotion recognition model was proposed to evaluate the classroom teaching effect [41]. The authors trained the CNN model and used it to predict the students' emotional states using the FER-2013 dataset. The proposed model demonstrates high accuracy and robustness in the detection and can evaluate the students' performance in the classroom in real-time so that teachers can get immediate feedback. However, the authors stated that some misclassifications still occur, such as those who wore spectacles or had bushy whiskers were predicted to be angry.

Moreover, a model to recognize learning based on CNNs and transfer learning was proposed [39]. The basic emotional model's transfer learning using the FER-2013 showed the significance of data complexity in the deep learning model's design process, achieving an accuracy rate of 84.59%. Nonetheless, the study focused only on demonstrating emotional data in a laboratory setting without appraising uncommon circumstances that could happen in the classroom environment.

In a previous study by Lasri et al., a model was proposed to integrate emotion recognition in education based on CNN [40]. Haar cascade was used in face detection, while CNN was used in emotion recognition and normalization using the FER-2013 dataset. The proposed model achieved an accuracy of 70%. Although the model effectively detects happy and surprising emotions, the emotion of fear is poorly detected, as it confuses the fear emotion with sad emotion.

Recently, the combination of CNN and the geometric feature-based method has been used to enhance the performance of the models [55]. The proposed model can monitor the students' faces through web cameras, and their facial expressions will be translated into learner engagement levels. Two tests have been carried out, and 90% of the CNN models have achieved an average accuracy of 95% for most subjects. The authors found that under certain circumstances, such as people having obstructions on their faces or moving their heads significantly, it was difficult to predict the levels of learning engagement.

Apart from that, genetic algorithms were also used to optimize CNN's hyperparameters to determine an individual's affective state in the previous study [53]. The results from the study demonstrate that the genetic algorithm enhances accuracy compared with CNN, which used the other machine learning algorithms and trial and error. The authors stated that the study results could be enhanced if the database is filtered and undergo a class distribution balancing.

*5.6.2. Long Short-Term Memory (LSTM).* LSTM networks are a special kind of RNN with a distinct and complex neural cell structure. They are specifically crafted to overcome the RNN's long-term dependence by using short-term memory. Despite the different structures of the repeating modules, the LSTM is still structured like a chain.

FER based on LSTM was previously proposed for video sequences because long-range context modelling can enhance the accuracy in analyzing emotion. There are several advantages of the LSTM model compared to standalone approaches for modelling sequential images. When integrating with other models, LSTM is straightforward in its end-to-end fine-tuning. The LSTM can also handle both variable-length and fixed-length inputs or outputs [110]. Furthermore, retaining the former cell unit information and updating the node information to the existing cell state value allows LSTM to learn the long-distance information in the dataset and pick the information generated by forgetting irrelevant information [111].

Recent studies have shown that LSTM is successfully applied in the FER field. A model that combined temporal convolution, bidirectional LSTM, and attention mechanism was proposed to determine the degree of engagement in online learning [47]. 60% of top-1 accuracy was achieved for four classifications using DAiSEE.

Next, a study that used the LSTM model to detect academic emotions such as boredom and frustration was conducted [52]. The findings indicated that, despite the higher accuracy of the FaceNet embeddings model, the facial landmark points model is more efficient at distinguishing between incidences of occurrence and nonoccurrence of boredom and frustration.

*5.6.3. Convolutional LSTM (ConvLSTM).* ConvLSTM is created as an extended and modified version of LSTM because of the limitations of LSTM [112]. It has an LSTM-like structure and can be applied for modelling long-term dependencies in either the spectral domain or time domain. ConvLSTM can be used to combine a CNN's capability in local data extraction with the ability of a RNN to use temporal context. The convolutional layers are used for feature extraction, while the transitions in image sequences are captured by LSTM layers.

Convolutional structures are available at both the input-to-state and state-to-state transitions in ConvLSTM. The spatial correlation is not taken into consideration by the conventional, fully connected LSTM. Since ConvLSTM is able to model spatiotemporal relationships, the convolution operator replaced the matrix multiplication in the LSTM formula. Moreover, the inputs and the past states of its local neighbours determine the future state of a certain state. This allows the model to capture long-term temporal relationships. Hence, ConvLSTM takes advantage of LSTM's ability

to handle temporal information and CNN's ability to handle spatial information.

A deep learning approach based on ConvLSTM was proposed for affective analysis [113]. The study results indicated that the proposed approach outperformed the conventional baseline and achieved the state-of-the-art system. Furthermore, a study conducted by Miyoshi et al. used an enhanced ConvLSTM to automatically recognize facial emotions from videos. Their findings indicated that the proposed technique achieved an accuracy of 49.26% for the eNTERFACE05 dataset and 95.72% for the CK+ dataset.

*5.7. Evaluation Method of FER.* FER is typically evaluated using metrics such as accuracy, precision, recall, and F1 score, depending on the specific task and dataset [15]. The evaluation metric is crucial in the training phase, and the selection is vital in distinguishing and obtaining the optimal classifier. Evaluation metrics, including accuracy, precision, recall, and F1 score, will be described in this section.

*5.7.1. Accuracy.* Accuracy is the number of data instances that are accurately classified over the sum of data instances. The issue with using accuracy as the primary performance metric is that it may not be reliable if there is a severe class imbalance. The following is the formula for accuracy:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}. \quad (1)$$

*5.7.2. Precision.* Precision is the classifier's ability to avoid labeling an instance positive that is negative and is also interpreted as the ratio of TP to the total of TP and FP for every class. Precision is a very useful metric that conveys more information compared to accuracy. In FER, precision is the proportion of automatic annotations for a specific action unit (AU) $i$ that are accurately identified by the model [114]. The following is the formula for precision:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}. \quad (2)$$

*5.7.3. Recall.* Recall is the classifier's ability to search all positive instances correctly. It is also interpreted as the fraction of TP to the total of TP and FN. This metric ranges from 0 to 1, with 1 being the best value. In FER, recall refers to the proportion of images containing a specific action unit (AU) $i$ that the model correctly identifies, calculated as the number of correct AU $i$ recognitions over the total number of images with AU $i$ [114]. The formula for recall is as follows:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \quad (3)$$

*5.7.4. F1-Score.* The F1-score is a weighted average between precision and recall, with 0.0 representing the worst and 1.0 representing the best score. This illustrates the importance of missing predicted classified emotions, which is a key factor based on weighted recall. The formula for the F1-score is as follows:

$$\text{F}_1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \quad (4)$$

*5.8. Critical Review.* In the previous study, Dewan et al. proposed a deep learning approach by using learners' facial expressions to detect their engagement [35]. Nonlinear correlation in the extracted features was captured using the KPCA, while the person-independent edge feature extraction was performed using the LDP for the multiple facial expressions. The experiments conducted on the DAiSEE demonstrated that the detection of two-level engagement achieved a better accuracy (90.89%) than the three-level engagement (87.25%). Nevertheless, the direct correlation between engagement and actual task performance is unknown.

In 2019, Bian et al. established a spontaneous facial expression dataset in the online learning environment named OL-SFED [46]. The dataset consists of five emotions, including confusion, distraction, enjoyment, fatigue, and neutral. In their studies, a CNN-based algorithm was applied and achieved an accuracy of 91.6%. Furthermore, it has been demonstrated through a comparison of pre- and postadoption assessment indicators that the method may significantly boost inference performance. However, the sample number of the dataset used in this study is limited.

In 2020, Leong conducted a study using the DAiSEE that only focused on detecting negative emotions such as boredom and frustration [52]. The findings showed that the facial landmark points model is more efficient at distinguishing between incidences of occurrence and nonoccurrence of boredom and frustration, even though the FaceNet embedding model's detection accuracy is higher. In my opinion, the authors should include positive emotions such as engagement so that educators can maintain students' interest and motivation.

In 2020, Rao and Rao proposed a hybrid CNN model to detect the cognitive state of the learner [61]. Datasets such as DAiSEE, JAFFE, and CK+ were used. Besides that, the classifier used in their study is SVM. The model has achieved 53.4%, 71.4%, and 99.95% for DAiSEE, JAFFE, and CK+, respectively. However, the "frustration" emotion's recognition rate is very low using DAiSEE since most of the images with a "frustration" cognitive state are predicted as "confusion".

Most of the existing similar studies that used academic emotion datasets are not real-time systems. Even though Huang et al. had built a real-time emotion recognition system using the DAiSEE, the accuracy rate is quite low, which is only 60% [47]. Furthermore, two studies that used the DAiSEE did not specifically classify student emotions. One of them only focuses on negative emotions, while another study grouped the emotions from the DAiSEE into two or three engagement levels. The previous studies that used the academic emotion datasets are summarized in Table 6 to highlight the gaps in previous studies.

Table 6: Selected publications that used academic emotion dataset.

| Reference | Emotion classifier | Type of emotions | Feature extraction | Results | Limitation | Real time |
|---|---|---|---|---|---|---|
| [97] | Long-term recurrent convolutional network | Boredom, confusion, engagement, and frustration | N/A | Top-1 accuracy: 94.6% | Low illumination and lack of frontal pose | No |
| [35] | Deep belief network (DBN) | Two-level engagement and three-level engagement | Viola-Jones algorithm | Accuracy:<br>(i) Two-level: 90.89%<br>(ii) Three-level: 87.25% | Unknown direct correlation between engagement and actual task performance | No |
| [46] | CNN | Confusion, distraction, enjoyment, fatigue, neutral | VGG16 | Accuracy: 91.60% | Limited sample number | No |
| [47] | BERN | Boredom, confusion, engagement, and frustration | OpenFace | Top-1 accuracy: 60% for four classification model | Requires a large amount of training data and a long training time | Yes |
| [52] | LSTM | Boredom and frustration | MTCNN library—face detection & cropping | Accuracy compared to the EmotioNet model:<br>(i) Boredom: +16.26%<br>(ii) Frustration -2.42% | (i) Decreased accuracy for frustration emotion<br>(ii) Involved negative emotions only | No |
| [61] | CNN | Boredom, confusion, engagement, and frustration | CNN+pose estimator | Accuracy: 53.4% | Show a low recognition rate for the frustration when using the DAiSEE dataset | Yes |
| [64] | Deep facial spatiotemporal network (DFSTN) | Engagement level: very low, low, high, and very high | MTCNN | Accuracy: 58.84% | (i) Low detection accuracy<br>(ii) Data deficiencies and data imbalances | No |

## 6. Challenges and Limitation

In recent years, FER in online learning has become an increasingly active research field. A variety of works have been done that have demonstrated remarkable outcomes and precisely classified emotions. The precision of classified emotions is likely being evaluated in relation to some preestablished benchmark or baseline, such as the performance of other emotion classifiers or the accuracy of human annotators. For example, in a study by Huang et al. in 2019, the performance of their model was compared to the state-of-the-art approaches and benchmarks on the dataset. Similarly, in a study by Mohan et al. in 2021, they tested their method on five benchmarking datasets and did a comparative evaluation of their model with 21 state-of-the-art methods. Despite the fact that deep learning-based FER techniques have been successful in experimental assessments, there are still a number of challenges and issues that need to be investigated further. In this section, the challenges and limitations will be discussed.

Firstly, there are currently relatively few datasets that can be accessed online that are applicable for the purpose of FER in the context of online learning. Nevertheless, the significance of the academic emotion dataset has recently come to light. Researchers are devoting an increasing amount of attention to the process of producing these types of datasets and making them publicly available. According to findings from previous research, it was difficult to establish a connection between certain facial emotions and learning tasks such as attending an online lecture, watching online video tutorials, writing, and reading [18].

Next, there is a requirement for a large amount of memory for deep learning models, and it is also time-consuming in model training [115]. Hence, the deep learning model is not well-suited for deployment on platforms with limited resources because of its high memory and sophisticated computing requirements. In order to obtain models that can be executed quickly without loss of accuracy, it is necessary to investigate methods for reducing the complexity.

Furthermore, the algorithms are developed for data exploitation and extrapolate stereotypical traits, which precludes them from considering exceptional cases and uncommon configurations [116]. According to cutting-edge theoretical perspectives, emotions are a nuanced and dynamic phenomenon that vary along many parameters that have not yet been completely formalized in theory.

Moreover, recent development in machine learning technologies, particularly deep learning like CNN, requires larger data sets than currently available [18]. The process of gathering and evaluating behavioural data in naturalistic settings is in and of itself a challenging task. Therefore, additional efforts are required to solve the open challenges related to the restrictions of the real-world learning environment.

## 7. Conclusion

This paper presents a systematic literature review of multiclass student emotion classification in online learning. In this paper, the scientific literature of the past five years was systematically searched to identify the type of FER approach used, the algorithm that was used as an emotion classifier, and the datasets used in the previous studies. It can be concluded that deep learning algorithm, such as CNN, is applied more frequently than other algorithms. On the other hand, SVM, random forest, decision tree, and KNN are examples of conventional machine learning algorithms used to classify facial emotions.

Based on the findings of this study, the deep learning approach is the most frequently adopted approach in classifying student emotion during online learning for the student FER systems in recent years. Furthermore, FER-2013 is the most commonly used FER dataset in FER studies, while DAiSEE is the most used academic emotion dataset. Moreover, support vector machine (SVM) is the conventional learning classifier that is widely used in the FER systems, while convolutional neural network (CNN) is the most frequently used deep learning classifier, followed by LSTM. Next, it was found that the number of real-time FER systems is less than the number of non-real-time FER systems. Finally, the top-1 accuracy of 94.6% was achieved by the long-term recurrent convolutional network on the academic emotion dataset in previous studies [97]. The limitation is that it has low illumination and a lack of frontal pose.

In conclusion, emotion recognition has come a long way over the years, with a significant number of approaches having been established. This has led to the emergence of new research issues, opportunities, and challenges, and as a result, the field has advanced one step in both the recognition and knowledge of emotions.

## 8. Future Directions

Although numerous studies on FER have been carried out previously, FER's performance has dramatically improved in recent years by combining deep-learning algorithms. Future studies must focus on developing annotation standards for labeling the benchmarking datasets. Academic-based affective states, such as boredom, frustration, and engagement, are more challenging to measure than the commonly investigated domains of emotion recognition [117]. By personalizing the datasets, specifically trained models can be created to recognize the students' unique academic emotions. This could lead to better insights into students' emotions and more personalized feedback and support, hence enhancing their learning outcomes.

Moreover, emotions can vary throughout time and are influenced by various internal and external circumstances. Long-term monitoring of student emotions can provide insightful information about their emotional states and assist in detecting patterns and trends. Future systems could explore approaches for long-term monitoring of facial emotions, such as wearable sensors or continuous video recording.

Besides that, several different ethical concerns are raised by emotion recognition systems, including privacy, consent, and potential biases [118]. These concerns could be addressed in future systems through transparent data collection and processing, informed consent, and unbiased algorithms.

In conclusion, the potential future directions for student FER systems are numerous and exciting. Advancements in dataset personalization, integration with other technologies, and ethical considerations hold great promise for improving students' learning experiences and outcomes. It is crucial to thoroughly consider the potential consequences of these advancements and implement them responsibly and ethically. This will allow us to utilize the potential of FER systems to create more engaging, personalized, and effective learning environments for students.

## Data Availability

The data supporting this systematic review are from previously reported studies and datasets, which have been cited.

## Conflicts of Interest

The authors declare that they have no conflicts of interest.

## Acknowledgments

The corresponding author is supported by a research grant from the Ministry of Higher Education, Malaysia (Fundamental Research Grant Scheme (FRGS), Dana Penyelidikan, Kementerian Pengajian Tinggi, FRGS/1/2019/ICT02/UMS/01/1). The APC is funded by Universiti Malaysia Sabah.

## References

[1] L. F. Barrett, R. Adolphs, S. Marsella, A. M. Martinez, and S. D. Pollak, "Emotional expressions reconsidered: challenges to inferring emotion from human facial movements," *Psychological Science in the Public Interest*, vol. 20, no. 1, pp. 1–68, 2019.

[2] L. Moustakas and D. Robrade, "The challenges and realities of e-learning during COVID-19: the case of university sport and physical education," *Challenges*, vol. 13, no. 1, p. 9, 2022.

[3] G. Marinoni, H. V. Land, and T. Jensen, "The Impact of Covid-19 on Higher Education around the World: IAU Global Survey Report," *International Association of Universities*, vol. 23, pp. 1–17, 2020.

[4] M. A. Almaiah, A. Al-Khasawneh, and A. Althunibat, "Exploring the critical challenges and factors influencing the e-learning system usage during COVID-19 pandemic," *Education and Information Technologies*, vol. 25, no. 6, pp. 5261–5280, 2020.

[5] L. B. Krithika and G. G. Lakshmi Priya, "Student emotion recognition system (SERS) for e-learning improvement based on learner concentration metric," *Procedia Computer Science*, vol. 85, pp. 767–776, 2016.

[6] S. Wibawanto and K. C. Kirana, "Recognition of student emotion based on matrix-1 median fisher's face and backpropagation algorithm," in *2017 International Conference on Electrical Engineering and Informatics (ICELTICs)*, pp. 103–108, Banda Aceh, Indonesia, 2017.

[7] W. Wang, K. Xu, H. Niu, and X. Miao, "Emotion recognition of students based on facial expressions in online education based on the perspective of computer simulation," *Complexity*, vol. 2020, Article ID 4065207, 9 pages, 2020.

[8] J. Camacho-Morles, G. R. Slemp, R. Pekrun, K. Loderer, H. Hou, and L. G. Oades, "Activity achievement emotions and academic performance: a meta-analysis," *Educational Psychology Review*, vol. 33, no. 3, pp. 1051–1095, 2021.

[9] M. Mukhopadhyay, S. Pal, A. Nayyar, P. K. D. Pramanik, N. Dasgupta, and P. Choudhury, "Facial emotion detection to assess learner's state of mind in an online learning system," in *Proceedings of the 2020 5th International Conference on Intelligent Information Technology*, pp. 107–115, Hanoi, Viet Nam, 2020.

[10] R. Pekrun and L. Linnenbrink-Garcia, "Academic emotions and student engagement," in *Handbook of Research on Student Engagement*, Springer, 2012.

[11] B. L. Fredrickson, "Positive emotions broaden and build," in *Advances in Experimental Social Psychology*, pp. 1–53, Elsevier, 2013.

[12] L. Rothkrantz, "New didactic models for MOOCs," in *Proceedings of the 9th International Conference on Computer Supported Education*, vol. 1, pp. 505–512, Porto, Portugal, 2017.

[13] H. Al-Samarraie, "A scoping review of videoconferencing systems in higher education," *International Review of Research in Open and Distance Learning*, vol. 20, no. 3, 2019.

[14] G. Siemens, "Massive open online courses: innovation in education," in *Athabasca: Commonwealth of Learning*, Athabasca University, 2013.

[15] B. C. Ko, "A brief review of facial emotion recognition based on visual information," *Sensors*, vol. 18, no. 2, p. 401, 2018.

[16] E. Sariyanidi, H. Gunes, and A. Cavallaro, "Automatic analysis of facial affect: a survey of registration, representation, and recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 6, pp. 1113–1133, 2015.

[17] P. Tarnowski, M. Kołodziej, A. Majkowski, and R. J. Rak, "Emotion recognition using facial expressions," *Procedia Computer Science*, vol. 108, pp. 1175–1184, 2017.

[18] M. A. A. Dewan, M. Murshed, and F. Lin, "Engagement detection in online learning: a review," *Smart Learning Environments*, vol. 6, no. 1, 2019.

[19] M. Soltani, H. Zarzour, and M. C. Babahenini, "Facial Emotion Detection in Massive Open Online Courses," in *Advances in Intelligent Systems and Computing*, pp. 277–289, Springer, 2018.

[20] I. M. Revina and W. R. S. Emmanuel, "A Survey on human face expression recognition techniques," *Journal of King Saud University - Computer and Information Sciences*, vol. 33, no. 6, pp. 619–628, 2021.

[21] N. P. N. Sreedharan, B. Ganesan, R. Raveendran, P. Sarala, B. Dennis, and R. Rajakumar Boothalingam, "Grey Wolf optimisation-based feature selection and classification for facial emotion recognition," *IET Biometrics*, vol. 7, no. 5, pp. 490–499, 2018.

[22] K. Patel, D. Mehta, C. Mistry et al., "Facial sentiment analysis using AI techniques: state-of-the-art, taxonomies, and challenges," *IEEE Access*, vol. 8, pp. 90495–90519, 2020.

[23] H. Sadeghi and A. A. Raie, "Human vision inspired feature extraction for facial expression recognition," *Multimedia Tools and Applications*, vol. 78, no. 21, pp. 30335–30353, 2019.

[24] R. S. Jadhav and P. Ghadekar, "Content Based Facial Emotion Recognition Model Using Machine Learning Algorithm," in *2018 International Conference on Advanced Computation*

*and Telecommunication (ICACAT)*, pp. 1–5, Bhopal, India, 2018.

[25] A. Mollahosseini, D. Chan, and M. H. Mahoor, "Going Deeper in Facial Expression Recognition Using Deep Neural Networks," in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1–10, Lake Placid, NY, USA, 2016.

[26] W. Mellouk and W. Handouzi, "Facial emotion recognition using deep learning: Review and insights," *Procedia Computer Science*, vol. 175, pp. 689–694, 2020.

[27] H. Dino, M. B. Abdulrazzaq, S. R. Zeebaree et al., "Facial expression recognition based on hybrid feature extraction techniques with different classifiers," *TEST Engineering & Management*, vol. 83, pp. 22319–22329, 2020.

[28] B. Kitchenham and S. Charters, "Guidelines for Performing Systematic Literature Reviews in Software Engineering," EBSE Technical Report EBSE-2007-01, 2007.

[29] S. Li and W. Deng, "Deep facial expression recognition: a survey," *IEEE Transactions on Affective Computing*, vol. 13, no. 3, pp. 1195–1215, 2022.

[30] G. Li and Y. Wang, "Research on leamer's emotion recognition for intelligent education system," in *2018 IEEE 3rd Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)*, pp. 754–758, Chongqing, China, 2018.

[31] O. El Hammoumi, F. Benmarrakchi, N. Ouherrou, J. El Kafi, and A. El Hore, "Emotion recognition in e-learning systems," in *2018 6th International Conference on Multimedia Computing and Systems (ICMCS)*, Rabat, Morocco, 2018.

[32] C. Ma, C. Sun, D. Song, X. Li, and H. Xu, "A deep learning approach for online learning emotion recognition," in *2018 13th International Conference on Computer Science & Education (ICCSE)*, pp. 1–5, Colombo, Sri Lanka, 2018.

[33] D. Yang, A. Alsadoon, P. W. C. Prasad, A. K. Singh, and A. Elchouemi, "An emotion recognition model based on facial recognition in virtual learning environment," *Procedia Computer Science*, vol. 125, pp. 2–10, 2018.

[34] K. Candra Kirana, S. Wibawanto, and H. Wahyu Herwanto, "Facial emotion recognition based on Viola-Jones algorithm in the learning environment," in *2018 International Seminar on Application for Technology of Information and Communication*, pp. 406–410, Semarang, Indonesia, 2018.

[35] M. A. A. Dewan, F. Lin, D. Wen, M. Murshed, and Z. Uddin, "A deep learning approach to detecting engagement of online learners," in *2018 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI),*, pp. 1895–1902, Guangzhou, China, 2018.

[36] A. Sharma and V. Mansotra, "Deep learning based student emotion recognition from facial expressions in classrooms," *International Journal of Engineering and Advanced Technology*, vol. 8, no. 6, pp. 4691–4699, 2019.

[37] A. Sharma and V. Mansotra, "Classroom student emotions classification from facial expressions and speech signals using deep learning," *International Journal of Recent Technology and Engineering (IJRTE)*, vol. 8, no. 3, pp. 6675–6683, 2019.

[38] L. Mao, N. Wang, L. Wang, and Y. Chen, "Classroom micro-expression recognition algorithms based on multi-feature fusion," *IEEE Access*, vol. 7, pp. 64978–64983, 2019.

[39] J. C. Hung, K. C. Lin, and N. X. Lai, "Recognizing learning emotion based on convolutional neural networks and transfer learning," *Applied Soft Computing*, vol. 84, article 105724, 2019.

[40] I. Lasri, A. R. Solh, and M. El Belkacemi, "Facial Emotion Recognition of Students Using Convolutional Neural Network," in *2019 Third International Conference on Intelligent Computing in Data Sciences (ICDS)*, pp. 1–6, Marrakech, Morocco, 2019.

[41] J. Tang, X. Zhou, and J. Zheng, "Design of Intelligent classroom facial recognition based on deep learning," *Journal of Physics: Conference Series*, vol. 1168, article 022043, 2019.

[42] Z. Shi, Y. Zhang, C. Bian, and W. Lu, "Automatic academic confusion recognition in online learning based on facial expressions," in *2019 14th International Conference on Computer Science & Education (ICCSE)*, pp. 528–532, Toronto, ON, Canada, 2019.

[43] M. Healy, R. Donovan, P. Walsh, and H. Zheng, "A machine learning emotion detection platform to support affective well being," in *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 2694–2700, Madrid, Spain, 2019.

[44] Y. Liang, "Intelligent emotion evaluation method of classroom teaching based on expression recognition," *International Journal of Emerging Technologies in Learning (iJET)International Journal of Emerging Technologies in Learning*, vol. 14, no. 4, pp. 127–141, 2019.

[45] S. Dash, M. A. Akber Dewan, M. Murshed, F. Lin, M. Abdullah-Al-Wadud, and A. Das, "A Two-Stage Algorithm for Engagement Detection in Online Learning," in *2019 International Conference on Sustainable Technologies for Industry 4.0 (STI)*, pp. 1–4, Dhaka, Bangladesh, 2019.

[46] C. Bian, Y. Zhang, F. Yang, W. Bi, and W. Lu, "Spontaneous facial expression database for academic emotion inference in online learning," *IET Computer Vision*, vol. 13, no. 3, pp. 329–337, 2019.

[47] T. Huang, Y. Mei, H. Zhang, S. Liu, and H. Yang, "Fine-grained engagement recognition in online learning environment," in *2019 IEEE 9th International Conference on Electronics Information and Emergency Communication (ICEIEC)*, pp. 338–341, Beijing, China, 2019.

[48] T. S. Ashwin and R. M. R. Guddeti, "Automatic detection of students' affective states in classroom environment using hybrid convolutional neural networks," *Education and Information Technologies*, vol. 25, no. 2, pp. 1387–1415, 2020.

[49] O. Mohamad Nezami, M. Dras, L. Hamey, D. Richards, S. Wan, and C. Paris, "Automatic recognition of student engagement using deep learning and facial expression," in *Machine Learning and Knowledge Discovery in Databases*, vol. 11908 of Lecture Notes in Computer Science, , pp. 273–289, Springer, 2020.

[50] T. Alrayassi and S. Shilbayeh, "Mood detection using facial recognition technology application in higher education institution," *International Journal of Advanced Trends in Computer Science and Engineering*, vol. 9, no. 5, pp. 7618–7627, 2020.

[51] X. Y. Tang, W. Y. Peng, S. R. Liu, and J. W. Xiong, "Classroom teaching evaluation based on facial expression recognition," in *Proceedings of the 2020 9th International Conference on Educational and Information Technology*, pp. 62–67, Oxford, United Kingdom, 2020.

[52] F. H. Leong, "Deep learning of facial embeddings and facial landmark points for the detection of academic emotions," in *Proceedings of the 5th International Conference on Information and Education Innovations*, pp. 111–116, London, United Kingdom, 2020.

[53] R. Zatarain Cabada, H. Rodriguez Rangel, M. L. Barron Estrada, and H. M. Cardenas Lopez, "Hyperparameter optimization in CNN for learning-centered emotion recognition for intelligent tutoring systems," *Soft Computing*, vol. 24, no. 10, pp. 7593–7602, 2020.

[54] X. Zhu and Z. Chen, "Dual-modality spatiotemporal feature learning for spontaneous facial expression recognition in e-learning using hybrid deep neural network," *The Visual Computer*, vol. 36, no. 4, pp. 743–755, 2020.

[55] A. Dubbaka and A. Gopalan, "Detecting learner engagement in MOOCs using automatic facial expression recognition," in *2020 IEEE Global Engineering Education Conference (EDUCON)*, pp. 447–456, Porto, Portugal, 2020.

[56] A. Pise, H. Vadapalli, and I. Sanders, "Facial emotion recognition using temporal relational network: an application to E-learning," *Multimedia Tools and Applications*, vol. 81, no. 19, pp. 26633–26653, 2022.

[57] N. Sabri, "Student emotion estimation based on facial application in e-learning during COVID-19 pandemic," *International Journal of Advanced Trends in Computer Science and Engineering*, vol. 9, no. 1.4, pp. 576–582, 2020.

[58] S. Kumar, D. Varshney, G. Dhawan, and H. Jalutharia, "Analysing the effective psychological state of students using facial features," in *2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS)*, pp. 648–653, Madurai, India, 2020.

[59] M. Murugappan, V. Maruthapillai, W. Khariunizam, A. M. Mutawa, S. Sruthi, and C. W. Yean, "Virtual markers based facial emotion recognition using ELM and PNN classifiers," in *2020 16th IEEE International Colloquium on Signal Processing & Its Applications (CSPA)*, pp. 261–265, Langkawi, Malaysia, 2020.

[60] M. Murugappan, A. M. Mutawa, S. Sruthi et al., "Facial Expression Classification Using KNN and Decision Tree Classifiers," in *2020 4th International Conference on Computer, Communication and Signal Processing (ICCCSP)*, pp. 1–6, Chennai, India, 2020.

[61] K. P. Rao and M. C. S. Rao, "Recognition of learners' cognitive states using facial expressions in e-learning environments," *Journal of University of Shanghai for Science and Technology*, vol. 22, no. 12, pp. 93–103, 2020.

[62] D. H. Hingu, "Facial expression analysis for emotion and behavior of online learner and framework for content adaptation," *International Research Journal of Engineering and Technology(IRJET)*, vol. 7, no. 7, 2020.

[63] B. E. Zakka and H. Vadapalli, "Estimating Student Learning Affect Using Facial Emotions," in *2020 2nd International Multidisciplinary Information Technology and Engineering Conference (IMITEC)*, pp. 1–6, Kimberley, South Africa, 2020.

[64] J. Liao, Y. Liang, and J. Pan, "Deep facial spatiotemporal network for engagement prediction in online learning," *Applied Intelligence*, vol. 51, no. 10, pp. 6609–6621, 2021.

[65] S. C. Siam, A. Faisal, N. Mahrab, A. B. Haque, and M. N. I. Suvon, "Automated Student Review System with Computer Vision and Convolutional Neural Network," in *2021 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS)*, pp. 493–497, Greater Noida, India, 2021.

[66] Q. Li, Y. Q. Liu, Y. Q. Peng et al., "Real-Time Facial Emotion Recognition Using Lightweight Convolution Neural Network," *Journal of Physics: Conference Series*, vol. 1827, no. 1, article 012130, 2022.

[67] K. Mohan, A. Seal, O. Krejcar, and A. Yazidi, "Facial expression recognition using local gravitational force descriptor-based deep convolution neural networks," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–12, 2021.

[68] K. Mohan, A. Seal, O. Krejcar, and A. Yazidi, "FER-net: facial expression recognition using deep neural net," *Neural Computing and Applications*, vol. 33, no. 15, pp. 9125–9136, 2021.

[69] K. Mohan and A. Seal, "Deception Detection on 'Bag-of-Lies': Integration of Multi-Modal Data Using Machine Learning Algorithms," in *Algorithms for Intelligent Systems*, pp. 445–456, Springer, 2021.

[70] M. Karnati, A. Seal, A. Yazidi, and O. Krejcar, "LieNet: a deep convolution neural network framework for detecting deception," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 14, no. 3, pp. 971–984, 2022.

[71] J. Shen, H. Yang, J. Li, and Z. Cheng, "Assessing Learning Engagement Based on Facial Expression Recognition in MOOC's Scenario," *Multimedia Systems*, vol. 28, no. 2, pp. 469–478, 2022.

[72] S. Gupta, P. Kumar, and R. K. Tekchandani, "Facial emotion recognition based real-time learner engagement detection system in online learning context using deep learning models," *Multimedia Tools and Applications*, vol. 82, no. 8, pp. 11365–11394, 2023.

[73] A. V. Savchenko, L. V. Savchenko, and I. Makarov, "Classifying emotions and engagement in online learning based on a single facial expression recognition neural network," *IEEE Transactions on Affective Computing*, vol. 13, no. 4, pp. 2132–2143, 2022.

[74] C. Hou, J. Ai, Y. Lin, C. Guan, J. Li, and W. Zhu, "Evaluation of online teaching quality based on facial expression recognition," *Future Internet*, vol. 14, no. 6, p. 177, 2022.

[75] Q. Yuan, "Research on classroom emotion recognition algorithm based on visual emotion classification," *Computational Intelligence and Neuroscience*, vol. 2022, Article ID 6453499, 10 pages, 2022.

[76] H. Wu, "Real time facial expression recognition for online lecture," *Wireless Communications and Mobile Computing*, vol. 2022, Article ID 9684264, 11 pages, 2022.

[77] S. Rajan, P. Chenniappan, S. Devaraj, and N. Madian, "Facial expression recognition techniques: a comprehensive survey," *IET Image Processing*, vol. 13, no. 7, pp. 1031–1040, 2019.

[78] Y. Li, S. Gong, and H. Liddell, "Support Vector Regression and Classification Based Multi-View Face Detection and Recognition," in *Proceedings Fourth IEEE International Conference on Automatic Face and Gesture Recognition (Cat. No. PR00580)*, pp. 300–305, Grenoble, France, 2000.

[79] S. L. Salzberg, "C4.5: Programs for Machine Learning by J. Ross Quinlan. Morgan Kaufmann Publishers, Inc., 1993," *Machine Learning*, vol. 16, no. 3, pp. 235–240, 1994.

[80] J. Naskath, G. Sivakamasundari, and A. A. S. Begum, "A study on different deep learning algorithms used in deep neural nets: MLP SOM and DBN," *Wireless Personal Communications*, vol. 128, no. 4, pp. 2913–2936, 2023.

[81] R. Wang, "AdaBoost for feature selection, classification and its relation with SVM, a review," *Physics Procedia*, vol. 25, pp. 800–807, 2012.

[82] J. Cao, S. Kwong, and R. Wang, "A noise-detection based AdaBoost algorithm for mislabeled data," *Pattern Recognition*, vol. 45, no. 12, pp. 4451–4465, 2012.

[83] J. Jia, Y. Xu, S. Zhang, and X. Xue, "The Facial Expression Recognition Method of Random Forest Based on Improved PCA Extracting Feature," in *2016 IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC)*, pp. 1–5, Hong Kong, 2016.

[84] Y. Wang, Y. Li, Y. Song, and X. Rong, "Facial expression recognition based on random forest and convolutional neural network," *Information*, vol. 10, no. 12, p. 375, 2019.

[85] T. U. Ahmed, S. Hossain, M. S. Hossain, R. Ul Islam, and K. Andersson, "Facial Expression Recognition Using Convolutional Neural Network with Data Augmentation," in *2019 Joint 8th International Conference on Informatics, Electronics & Vision (ICIEV) and 2019 3rd International Conference on Imaging, Vision & Pattern Recognition (icIVPR)*, pp. 336–341, Spokane, WA, USA, 2019.

[86] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017.

[87] A. Nguyen, J. Yosinski, and J. Clune, "Deep Neural Networks Are Easily Fooled: High Confidence Predictions for Unrecognizable Images," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 427–436, Boston, MA, USA, 2015.

[88] W. Pannakkong, K. Thiwa-Anont, K. Singthong, P. Parthanadee, and J. Buddhakulsomsiri, "Hyperparameter tuning of machine learning algorithms using response surface methodology: a case study of ANN, SVM, and DBN," *Mathematical Problems in Engineering*, vol. 2022, Article ID 8513719, 17 pages, 2022.

[89] L. Nwosu, H. Wang, J. Lu, I. Unwala, X. Yang, and T. Zhang, "Deep Convolutional Neural Network for Facial Expression Recognition Using Facial Parts," in *2017 IEEE 15th Intl Conf on Dependable, Autonomic and Secure Computing, 15th Intl Conf on Pervasive Intelligence and Computing, 3rd Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress(DASC/PiCom/DataCom/CyberSciTech)*, pp. 1318–1321, Orlando, FL, USA, 2017.

[90] N. Jain, S. Kumar, A. Kumar, P. Shamsolmoali, and M. Zareapoor, "Hybrid deep neural networks for face emotion recognition," *Pattern Recognition Letters*, vol. 115, pp. 101–106, 2018.

[91] P.-L. Carrier and A. Courville, "The Facial Expression Recognition 2013 (FER-2013) Dataset," Wolfram Data Repository, 2013.

[92] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, "Coding facial expressions with Gabor wavelets," in *Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition*, pp. 200–205, Nara, Japan, 1998.

[93] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended Cohn-Kanade dataset (CK+): a complete dataset for action unit and emotion-specified expression," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, pp. 94–101, San Francisco, CA, USA, 2010.

[94] D. Lundqvist, A. Flykt, and A. Ohman, "The Karolinska Directed Emotional Faces (KDEF)," in *CD ROM from Department of Clinical NeurosciencePsychology Section*, Karolinska Institutet, 1998.

[95] S. M. Mavadati, M. H. Mahoor, K. Bartlett, P. Trinh, and J. F. Cohn, "DISFA: a spontaneous facial action intensity database," *IEEE Transactions on Affective Computing*, vol. 4, no. 2, pp. 151–160, 2013.

[96] M. Mavadati, P. Sanger, and M. H. Mahoor, "Extended DISFA dataset: investigating posed and spontaneous facial expressions," in *2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1452–1459, Las Vegas, NV, USA, 2016.

[97] A. Gupta, A. D'Cunha, K. Awasthi, and V. Balasubramanian, "DAiSEE: towards user engagement recognition in the wild," vol. 14, no. 8, pp. 1–12, 2016, http://arxiv.org/abs/1609.01885.

[98] C. Rahmad, R. A. Asmara, D. R. H. Putra, I. Dharma, H. Darmono, and I. Muhiqqin, "Comparison of Viola-Jones Haar Cascade Classifier and Histogram of Oriented Gradients (HOG) for Face Detection," *IOP Conference Series: Materials Science and Engineering*, vol. 732, no. 1, article 012038, 2020.

[99] T. M. Effendi, H. B. Seta, and T. Wati, "The Combination of Viola-Jones and Eigen Faces Algorithm for Account Identification for Diploma," *Journal of Physics: Conference Series*, vol. 1196, article 012070, 2019.

[100] I. Talegaonkar, K. Joshi, S. Valunj, R. Kohok, and A. Kulkarni, "Real time facial expression recognition using deep learning," *SSRN Electronic Journal*, 2019.

[101] U. Ayvaz, H. Gürüler, and M. O. Devrim, "Use of facial emotion recognition in E-learning systems," *Information Technologies and Learning Tools*, vol. 60, no. 4, p. 95, 2017.

[102] Y. H. Liu, "Feature Extraction and Image Recognition with Convolutional Neural Networks," *Journal of Physics: Conference Series*, vol. 1087, article 062032, 2018.

[103] A. Rahim, N. Hossain, T. Wahid, and S. Azam, "Face recognition using local binary patterns (LBP)," *Global Journal of Computer Science and Technology*, vol. 13, no. 4, 2013.

[104] P. C. Vasanth and K. R. Nataraj, "Facial expression recognition using SVM classifier," *Indones. J. Electr. Eng. Informatics*, vol. 3, no. 1, 2015.

[105] S. Gupta, K. Verma, and N. Perveen, "Facial expression recognition system using facial characteristic points and ID3," *International Journal of Computer and Communication Technology*, vol. 3, no. 1, pp. 45–49, 2014.

[106] D. Canedo and A. J. R. Neves, "Facial expression recognition using computer vision: a systematic review," *Applied Sciences*, vol. 9, no. 21, p. 4678, 2019.

[107] Y. Huang, F. Chen, S. Lv, and X. Wang, "Facial expression recognition: a survey," *Symmetry*, vol. 11, no. 10, p. 1189, 2019.

[108] R. Yamashita, M. Nishio, R. K. G. Do, and K. Togashi, "Convolutional neural networks: an overview and application in radiology," *Insights Into Imaging*, vol. 9, no. 4, pp. 611–629, 2018.

[109] A. T. Lopes, E. de Aguiar, A. F. De Souza, and T. Oliveira-Santos, "Facial expression recognition with convolutional neural networks: coping with few data and the training sample order," *Pattern Recognition*, vol. 61, pp. 610–628, 2017.

[110] J. Donahue, L. A. Hendricks, M. Rohrbach et al., "Long-term recurrent convolutional networks for visual recognition and description," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 4, pp. 677–691, 2017.

[111] J. Yang, X. Huang, H. Wu, and X. Yang, "EEG-based emotion classification based on bidirectional long short-term memory network," *Procedia Computer Science*, vol. 174, pp. 491–504, 2020.

[112] X. Shi, Z. Chen, H. Wang, D. Y. Yeung, W. K. Wong, and W. C. Woo, "Convolutional LSTM network: a machine learning approach for precipitation nowcasting," *Advances in Neural Information Processing Systems*, vol. 28, 2015.

[113] J. Niu, S. Li, S. Mo, Y. Guo, and L. Wang, "Affective analysis for video frames using convLSTM network," in *2018 IEEE International Conference on Communications (ICC)*, pp. 1–6, Kansas City, MO, USA, 2018.

[114] C. F. Benitez-Quiroz, R. Srinivasan, and A. M. Martinez, "EmotioNet: An Accurate, Real-Time Algorithm for the Automatic Annotation of a Million Facial Expressions in the Wild," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5562–5570, Las Vegas, NV, USA, 2016.

[115] J. Gu, Z. Wang, J. Kuen et al., "Recent advances in convolutional neural networks," *Pattern Recognition*, vol. 77, pp. 354–377, 2018.

[116] A. Masson, G. Cazenave, J. Trombini, and M. Batt, "The current challenges of automatic recognition of facial expressions: a systematic review," *AI Communications*, vol. 33, no. 3-6, pp. 113–138, 2020.

[117] J. Whitehill, Z. Serpell, Y. C. Lin, A. Foster, and J. R. Movellan, "The faces of engagement: automatic recognition of student Engagement from facial expressions," *IEEE Transactions on Affective Computing*, vol. 5, no. 1, pp. 86–98, 2014.

[118] R. A. Waelen, "The struggle for recognition in the age of facial recognition technology," *AI and Ethics*, vol. 3, no. 1, pp. 215–222, 2023.