Chapter 2
Systematic Literature Review

Technology-enhanced learning is transforming education and reshaping educational institutions by integrating advanced digital tools to create more engaging, accessible, and effective learning experiences [1]. TEL encompasses a broad range of technologies, from interactive software and multimedia content to virtual reality and artificial intelligence. This approach not only enhances the learning experience but also aligns with modern students' expectations and comfort with technology, making learning more relevant and accessible in a digitally driven society.

The importance of TEL lies in its ability to adapt education to the needs of today's tech-savvy learners. With young people spending a significant portion of their day interacting with digital devices, they are accustomed to and comfortable with technology. By integrating TEL into educational systems, educators can meet students on familiar ground, using the same tools and platforms they are already accustomed to. TEL not only meets the current educational standard but also significantly improves learning outcomes by personalizing the learning experience, providing instant feedback, and allowing students to engage at their own pace.

An essential aspect of TEL is its ability to recognize and respond to students' emotional states, a key indicator of engagement levels. By identifying emotions like frustration, boredom, or enthusiasm, TEL systems can adapt content, offer additional support, or provide motivating challenges, thus maintaining an optimal level of engagement. Emotion recognition, powered by artificial intelligence, can offer insights into learners' needs and preferences, making it easier to provide a flexible and supportive learning environment.

Mulsemedia, which includes video, audio, interactive simulations, and animations, plays a crucial role in TEL by making learning more dynamic and engaging. Integrating multimedia content into educational platforms allows for a richer learning experience, catering to different learning styles and making complex concepts easier to understand [13]. With TEL, multimedia can be seamlessly incorporated into lessons, enabling more engaging and immersive learning experiences and helping students retain information better.

This chapter presents a systematic review of assessing learner emotions through FER, aiming to provide a comprehensive understanding of the practical development of FER approaches. The discussion delves into the essential components of DL techniques to fully capture the motivational aspects within the realm of FER. Additionally, the chapter explores the mulsemedia approach in learning environments to deliver mulsemedia-synchronized learning content, highlighting its significance and potential impact on enhancing the learning experience.

The remainder of this chapter is organized as follows: Section 2.1 presents the methodology used for selecting research articles. Section 2.2 describes the mulsemedia approach within a learning context to enhance the quality of learning experiences. Section 2.3 discusses learner engagement detection through FER techniques. Section 2.4 describes publicly available datasets related to FER. Section 2.5 provides a brief overview of FER in machine learning, covering pre-processing, feature extraction, and classification approaches. Section 2.6 explores FER models based on deep learning (DL). Section 2.7 explains eXplainable AI techniques for interpreting model prediction results. Section 2.8 provides the research gaps based on the systematic review analysis. Finally, Section 2.9 summarizes the key points discussed in this chapter.

2.1 Methodology of the survey

The Systematic Literature Review (SLR) [24] aimed to select papers that improve the quality of reporting, identify existing research gaps on specific research problems, and enhance the efficiency of the review process for both FER researchers and practitioners. These individuals are interested in addressing key research challenges related to analyzing learners' emotions in learning environments. Additionally, this study adhered to the PRISMA guidelines [24], an evidence-based set of recommendations designed to improve the quality and transparency of systematic reviews and meta-analyses. For this critical analysis, the SLR review protocol was initially prepared to collect literature, following guidelines published in 2007. The review protocol was then evaluated iteratively, as it was a recursive process. This protocol helped minimize potential publication bias. First, research questions (RQs) were formulated. Four research questions were established to guide this SLR:

RQ 1: *What are the most significant processes in FER using ML and DL techniques?*

RQ 2: *What challenges are associated with FER databases?*

RQ 3: *How many facial expressions are the main focus of the FER system?*

RQ 4: *What are the accuracy and limitations of existing FER studies when using ML and DL techniques?*

RQ 5: *How are universal facial expressions mapped in FER in a learning environment?*

The search string and domain list if frequently altered until the search results showed significant results for each identified domain found. Next, the publications were identified and chosen by searching available databases during paper selection. The paper extraction process is done through authors' details, publication types and year, and other details were asked in the research question. After this step, data synthesis was made to present an overview of the related studies published up to date. The review was conducted at the last stage by reporting and answering the search questions. The review should be thoroughly examined in adequate detail for researchers and practitioners to evaluate the comprehensive search. Also, the unfiltered research results have been stored for further analysis if there is required in the future.

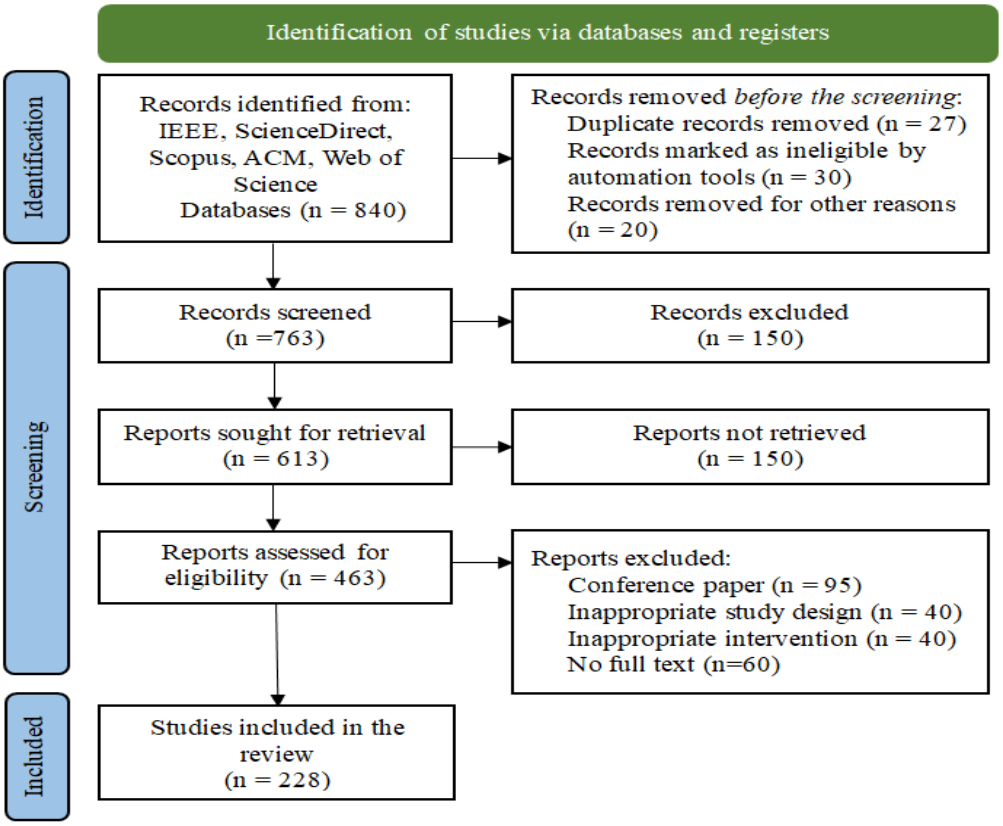

Figure 2.1 Screening process is illustrated using the PRISMA Flow Chart

The papers were retrieved from six different standard databases such as IEEE, Web of Science, Scopus, ACM, Springer, and Science Direct. The initial search string was "Facial expression recognition" AND "Facial emotion recognition". Open Access papers, journals, conference materials, and manuscripts were used in the search method. The final search string is as follows: (("Facial Expression Recognition" OR "Facial Emotion Recognition" OR "Facial Expression Analysis" OR "Facial Expression Recognition E-Learning") AND ("Emotion Recognition" OR "Emotion Detection" OR "E-Learning") AND ("machine learning in FER" OR "deep learning in FER")). There are many keywords available to search the FER manuscript, but search techniques have used limited keywords for selecting more appropriate papers. Also, the search string was additionally filtered with the years 2002 to 2023, open access papers and the language is English.

All selected papers are carefully validated to exclude the inappropriate papers for this review analysis, resulting in a total of 763 non-duplicated papers. First, 613 publications were selected after a rapid scan of all the titles and abstracts to weed out research that did not apply to the subject of this systematic review or did not match the selection criteria. The remaining studies that did not fit the selection criteria were then removed by carefully reading the entire texts of the remaining publications, leaving 463 papers. The following listed exclusion criteria (EC) were used to exclude irrelevant studies:

EC 1: Theoretical studies and improper FER results in conference papers

EC 2: Papers are not related computer science field.

EC 3: Survey and short communication papers

EC 4: Duplicate publication of multiple sources

EC 5: Older version of FER research techniques before 2002

Following the listed EC criteria, finally, 228 papers were left for further review. In order to answer the RQ accordingly, the data from the database was extracted and synthesized. The systematic review's search and exclusion procedures are summarized in Figure 2.1

2.2 Mulsemedia Approach in Learning Context

Mulsemedia refers to the integration of multiple sensory modalities, such as visual, auditory, tactile, and interactive elements, into learning content. In the field of education, mulsemedia plays a crucial role in enhancing learner engagement, understanding, and retention.

Recently, researchers have begun exploring the effects of mulsemedia by incorporating additional sensory stimuli targeting the olfactory, airflow, tactile, and gustatory senses. These additional stimuli may further enhance user experiences. Many studies have also examined the olfactory effect in conjunction with traditional media, aiming to improve information retention and learning performance. While the results of these studies demonstrate that mulsemedia positively impacts user satisfaction when engaging with such content, research on mulsemedia in learning contexts remains limited. Most studies focus primarily on entertainment, virtual reality technologies, and audio-book applications. As such, there is a growing interest in investigating the effects of mulsemedia within learning environments, particularly in science-related subjects.

Mayer [25] researched multimedia learning, examining how the human mind processes information when exposed to both words and images. This study concluded that the combination of words and images engages both auditory and visual senses, enhancing knowledge acquisition.

Ghergulescu and Muntean [26] discussed the latest trends in game-based e-learning assessment, with a specific focus on assessing learner motivation. They presented methods for gathering information on player/learner motivation, which include dialog-based interaction, game-play-based interaction, and the use of additional equipment. The authors explored how game-based learning can effectively motivate learners to engage actively in the learning process.

Andonova et al. [27] proposed practical recommendations for utilizing multisensory stimulation, including virtual reality and scent, to assist educators in developing effective teaching strategies. These strategies are aimed at enhancing various aspects of the learning experience, recall, and creativity within a typical learning environment. Their findings indicated that when traditional video content was combined with a coherent olfactory stimulus, it led to higher self-reported ratings of the perceived quality of the sensory experience. Furthermore, combining an olfactory stimulus with either VR or traditional video resulted in higher self-reported ratings of perceived immersion. In terms of recall, traditional video alone yielded the highest scores in a typical learning setting. However, both VR alone and in conjunction with an olfactory stimulus showed improvements in enhancing creativity.

Covaci et al. [28] conducted a study to explore the impact of quality degradation on the sense of presence in immersive VR applications. Additionally, they aimed to expand the capabilities of 360-degree technology by incorporating multisensory stimuli. The study involved 48 participants who experienced both 360-degree scenarios with and without multisensory content. These participants were randomly assigned to four conditions characterized by different encoding qualities (HD, FullHD, 2.5K, 4K). The findings revealed that the sense of presence was not influenced by streaming at a higher bitrate. However, an interesting trend emerged, indicating that the inclusion of multisensory content had a positive and significant impact on the sense of presence. This highlights the importance of multisensory technology in creating more immersive experiences in VR applications.

Tijou et al. [29] conducted a study to investigate the effects of olfaction on the learning, retention, and recall of complex 3D structures, such as organic molecules. The study was carried out in both desktop and immersive configurations. In the first case, students were asked to use head motions to rotate shapes with the mouse. In the second case, they asked subjects to sit in front of a large screen to examine the molecular structures. To facilitate the study, researchers used olfactory devices to emit relevant scents, specifically employing spearmint scent with D-Carvone organic compounds. The authors found that participants were actively engaged with these systems.

Cuturi et al. [30] presented multisensory learning to teach mathematical concepts, aiming to overcome difficulties in acquiring mathematical competence in conventional classroom settings. The initiative began with elementary school teachers' experiences in addressing learning difficulties among children, especially concerning arithmetic and geometric concepts. In addition, a questionnaire was administered to assess the feasibility of employing multisensory learning, particularly through haptic, audio, and visual stimuli, for teaching mathematical concepts.

Zou et al. [31] introduced a mulsemedia-enhanced VideoLan Player (VLC), designed to deliver educational video content with haptic, olfactory, and airflow effects. The authors showed that mulsemedia-based learning provides an enjoyable experience for students who are open to immersive learning in their curriculum. However, it is noteworthy that the study did not specifically focus on the learning efficiency of acquired knowledge. In this

experiment, students were asked to evaluate the learning experience involving various sensory media, resulting in effects that enhance overall learning experiences in the educational context.

Tal et al. [32] conducted a study to explore the effects of multi-sensorial media (mulsemedia) in learning among postgraduate students at Dublin City University in Ireland. The research analyzed how this innovative TEL method influenced learner experience, motivation, and learning outcomes. The findings revealed significant improvements in academic performance, learning experience, engagement, and motivation among students who engaged in mulsemedia-enhanced learning. Surveys indicated that over 80% of participants enjoyed using mulsemedia during learning and found it highly motivating. Furthermore, approximately 70% expressed a desire to continue experiencing mulsemedia-enhanced learning methods.

Alkasasbeh and Ghinea [33] conducted a study investigating the potential of olfactory media as an informational cue and its impact on learning performance and user QoE. They developed an olfactory-enhanced quiz (web-based) focused on four countries, incorporating different types of questions using text, images, audio, and olfactory media. Four scents related to the respective countries were used in the quiz. Sixty-four participants were involved in the experiment to assess this application. The results indicated that the use of olfactory media alongside traditional digital media significantly improved learner performance compared to scenarios without olfactory cues. Regarding user QoE, olfactory media had a positive influence, and participants expressed enthusiasm about engaging with enhanced olfactory applications in the future.

Tal et al. [34] explored innovative TEL methodologies, particularly mulsemedia, within the context of the NEWTON project. NEWTON, part of the Horizon 2020 European initiative, introduces groundbreaking TEL methodologies and tools into a pan-European STEM learning network platform. The study delved into mulsemedia's impact on teaching and learning STEM subjects, focusing on telecommunication and networking modules. Through case studies, the research analyzed how mulsemedia-enhanced education positively influences the learning process in these specialized areas. This study was done with 42 engineering students from two different countries and learners' satisfaction levels were

analyzed with QoE questionnaires. The findings highlighted a significant improvement in students' learning experiences and knowledge acquisition with mulsemedia-enhanced education.

Bi et al. [35] investigated the educational potential of mulsemedia and introduced a Dynamic Adaptive Streaming over HTTP (DASH)-based Multi-sensory Media Delivery Solution (DASHMS). This innovative solution enables the adaptive distribution of mulsemedia content based on various factors such as network conditions, device capabilities, and user preferences. The study conducted a real-life educational experiment involving 44 students at an Irish university to evaluate DASHMS. The evaluation focused on both learner satisfaction and the impact on learning outcomes. The results indicated a significant enhancement in user experience with adaptive multi-sensorial media delivery. While memory recall showed a statistically significant improvement, the overall impact on learning outcomes beyond memory recall was not as pronounced in the experiment.

Mesfin et al. [36] introduced an eye-tracking device and a heart rate monitor wristband to monitor users' eye gaze and heart rate during their interaction with mulsemedia. Following each video clip, participants were prompted to complete an on-screen questionnaire covering aspects of smell, sound, and haptic effects to assess their enjoyment and perception of the experiment. The analysis of eye gaze and heart rate data revealed a significant impact of cross-modally mapped multisensorial effects on users' QoE. Specifically, the results indicated that when olfactory content aligns cross-modally with visual content, users' visual attention tends to focus more on the corresponding visual feature. Additionally, crossmodally matched media led to an enhanced QoE compared to conditions where only video content was presented.

Raheel et al. [37] conducted a study to create a novel dataset of multimodal physiological signals aimed at recognizing emotions in response to multimedia content. In pursuit of this goal, four multimedia clips were carefully selected and synchronized with sensory devices including a fan, heater, olfaction dispenser, and haptic vest to enhance the experience with cold air, hot air, olfaction, and haptic effects, respectively. The researchers monitored physiological responses such as electroencephalogram (EEG), galvanic skin response (GSR), and photoplethysmography (PPG) to analyze human emotional reactions during exposure to mulsemedia content. Statistical analysis, including a t-test based on arousal and valence

scores, revealed that engaging more than two human senses elicited significantly distinct emotions. Furthermore, statistical tests on EEG, GSR, and PPG responses indicated a notable difference between traditional multimedia and mulsemedia content. The study achieved a classification accuracy of 85.18% for valence and 76.54% for arousal using a K-nearest neighbor classifier and a feature-level fusion strategy.

2.3 Learner Engagement Detection through FER

Learner engagement detection through FER is a compelling application of technology in the learning environment. By leveraging FER systems, educators can gain valuable insights into learners' engagement levels during e-learning. These systems analyze facial cues, such as smiles, frowns, raised eyebrows, and eye contact, to assess learners' attention, interest, and understanding. This information is crucial for instructors, as it allows them to tailor teaching strategies, adjust content delivery, and provide targeted interventions that enhance learner engagement and improve learning outcomes. Additionally, FER-based engagement detection fosters more interactive and personalized learning experiences, creating a conducive environment for effective learning and knowledge retention.

Sun et al. [38] introduced a CNN-based FER model to detect learners' emotions in an e-learning system. They evaluated the model's performance using datasets such as CK+, JAFFE, and NVIE. Additionally, they tested the designed e-learning system in a real-time environment.

Pise et al. [39] proposed a solution involving the implementation of a deep-learning-based facial image analysis model to estimate learning affect and reflect on the level of student engagement. The authors introduced the Temporal Relational Network (TRN) for identifying changes in emotions on students' faces during e-learning sessions. They observed that TRN sparsely samples individual frames and learns their causal relations, which is more efficient than sampling dense frames and convolving them. The framework considers both single-scale and multi-scale temporal relations to achieve its goal. Additionally, a Multi-Layer Perceptron (MLP) was tested as a baseline classifier. The proposed framework is end-to-end trainable for video-based FER and was tested on the open-source DISFA+ database. The TRN-based model exhibited a significant reduction in the length of the feature set, which

proved effective in recognizing expressions. It was observed that the multi-scale TRN achieved better accuracy (92.7%) than the single-scale TRN (89.4%) and MLP (86.6%).

Zhu and Chen [40] introduced a novel approach for recognizing facial expressions in e-learning using a dual-modality spatiotemporal feature representation learning with a hybrid deep neural network. In addition to facial expression class information, our study leverages representative expression states such as onset, apex, and offset of expressions for improved recognition. The hybrid deep neural network learns spatiotemporal geometrical feature representations and spatial-temporal appearance feature representations. These dual modalities feature fusion representations are then employed for facial expression recognition. Comprehensive experiments were conducted on two spontaneous micro-expression datasets (CAS(ME)2 and CASME II), demonstrating that the proposed method achieved higher recognition accuracy compared to state-of-the-art methods.

Savchenko et al. [41] presented a video facial processing pipeline for analyzing student engagement and emotions. The pipeline begins with face detection, tracking, and clustering techniques to extract sequences of faces for each student. Subsequently, a single efficient neural network is utilized to extract emotional features from each frame. This network undergoes pre-training for face identification and fine-tuning for facial expression recognition using a robust optimization technique developed specifically for this purpose. The study demonstrates that these facial features enable rapid simultaneous prediction of students' engagement levels (from disengaged to highly engaged), individual emotions (such as happiness or sadness), and group-level effects (positive, neutral, or negative).

Rao and Rao [42] introduced a hybrid-CNN model designed to recognize a learner's cognitive state by combining manually engineered features with features extracted from a convolutional neural network. The model's performance was compared with both manual feature extraction methods and CNN methods independently. This proposed method was trained and tested using the DAiSEE spontaneous database tailored for e-learning contexts, as well as established datasets like Japanese Female Facial Expression (JAFFE) and Extended Cohn-Kanade Dataset (CK+). The model achieved accuracy rates of 53.4%, 71.4%, and 99.95%, respectively.

Du, Crespo, and Martínez [43] introduced a heuristic multimodal real-time emotion recognition approach (HMR-TER) aimed at providing timely and relevant online feedback based on learners' vocal intonations and facial expressions to enhance their learning experience. They introduced hybrid validation dynamic analysis to address the overall lack of learner motivation in e-learning settings. The final results showed improvements in various metrics: the face detection ratio increased by 84.25%, hand gesture recognition improved by 92.70%, voice recognition accuracy reached 82.26%, emotional problems were reduced by 84.5%, and the efficiency of e-learning activities rose to 93.85%.

Zakka and Vadapalli [44] created a platform designed to offer real-time feedback to learners during online learning videos. This platform employs a CNN to detect, predict, and analyze learners' facial emotions, which are then mapped to their learning affect. The feedback generated aims to provide a meaningful assessment of the learner's comprehension level.

2.4 Facial Expression Dataset

This section provides some existing popular databases that are frequently used in FER in both the training and testing phases. The performance of the system is affected when training is not done with sufficient datasets. Most of the facial expression dataset contains six basic emotions plus neutral, which consists only of the frontal face with some challenges like pose variation and illumination. Table 2.1 shows some popular publicly available databases used in the FER system in recent decades, and it also contains recently released facial expression datasets. In general, FER datasets have been categorized under two conditions: (1) Spontaneous and (2) Posed.

Spontaneous datasets are captured in a natural way, presenting more realistic expressions, and are often considered genuine and authentic. Posed expressions, on the other hand, are deliberately created and controlled by subjects. Early research exploring facial expressions primarily used posed expressions, where respondents were instructed to exhibit or reproduce each basic six expressions, which expression may not always accurately identify participants' emotions on their faces [45]. Most FER systems are still trained using posed expressions along with spontaneous ones due to a lack of a sufficient count of spontaneous expressions in the available dataset. Figure 2.2 shows the sample facial expression of spontaneous vs. posed.

Figure 2.2 Posed vs. Spontaneous Facial Expression

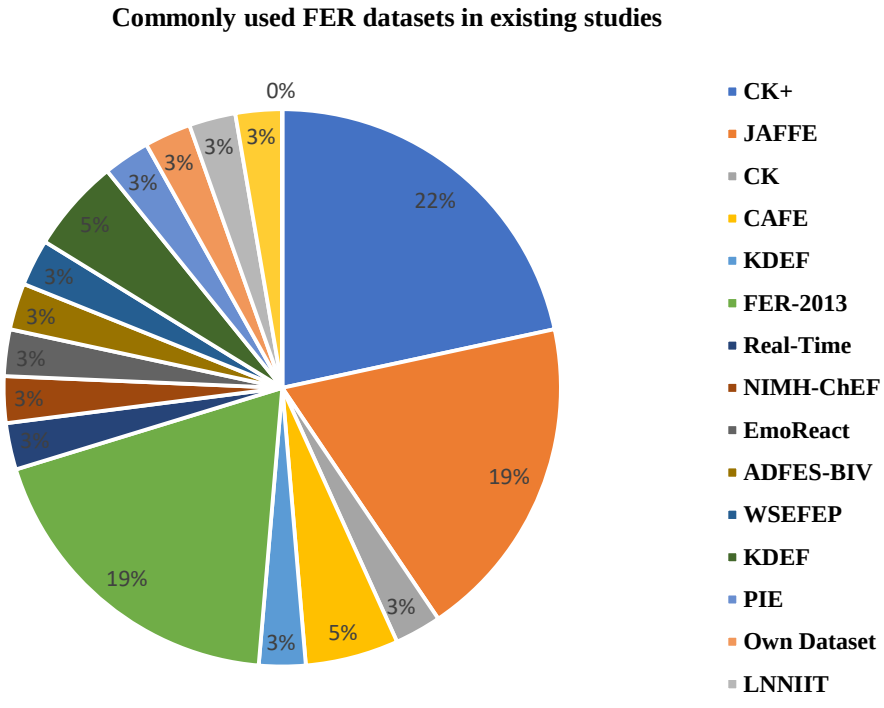

Figure 2.3 Most commonly used Facial Expression Datasets

Table 2.1 Publicly Available FER dataset

Name of the Database	Dataset	Subjects	No. of Expression	Condition	Age group	Posed/ Spontaneous	Created/ released year
JAFFE [46]	213 images	10	7	Lab	Adults	P	1998
KDEF [47]	4,9000 images	140	7	Lab	20-30	P	1998
CK+ [48]	593 images	123	7	Lab	Adults	P & S	2010
MMI [49]	740 images, 2,900 videos	32	7	Lab	Adults	P & S	2005/2010
Multi-PIE [50]	755,370 images	337	6	Lab	20-30	P	2010
Oulu-CASIA [51]	2,880 images	80	6	Lab	23-58	P	2011
FER-2013 [52]	35,887 images	N/A	7	Internet	20-60	P & S	2013
EmotioNet [53]	10,00,000 images	N/A	23	Internet	N/A	P & S	2016
AffectNet [54]	4,50,0000 images	N/A	7	Internet	N/A	S	2017
Ferv39k [55]	38,935 videos	N/A	7	Internet Videos	N/A	S	2021
PEDFE [56]	1458	56	6	Lab	20-30	P	2022
UIBVFED-Mask [57]	640 images	20	32	Lab	20-80	P	2023

*P=Posed; S=Spontaneous

JAFFE [46]: The JAFFE database consists of 213 photographic images capturing the posed facial expressions of 10 female Japanese individuals. Each image resolution is 256x256 pixels. It includes six primary emotions along with neutral expressions. This database is openly accessible for non-commercial research purposes.

KDEF [47]: The KDEF (Karolinska Directed Emotional Faces) consists of a total of 4900 images of facial expressions. It includes 17 subjects, 7 different facial expressions of 5 different angles.

CK+ [48]: The CK+ database comprises 593 sequences obtained from 123 subjects. The sequences do not have fixed lengths, and their durations range from 10 to 15 frames. It contains 6 primary expressions besides neutral along with facial landmark location. Out of 593 videos, only 309 were labeled as six basic emotions. It is also available to all kinds of researchers. The images in the database have a pixel resolution of 640 x 480 and 640 x 490 pixels, while their grey levels are represented in an 8-bit precision format.

MMI [49]: The Multimedia Interface (MMI) database includes 740 images and 2,900 videos from 32 subjects. A total of 213 images are labeled with six primary emotions. The pixel of the image size is 720 x 576 resolution.

Multi-PIE [50]: The Multi-PIE database contains 7,50,000 photography pictures of 337 subjects, which includes 15 viewpoints, 19 illumination conditions, and 5 different facial expressions.

Oulu-CASIA [51]: It contains six facial expressions from 80 subjects. The size of the frame is 320 x 420-pixel resolution of 25 frames per second. Also, the camera distance face is about 60 cm.

FER-2013 [52]: The FER-2013 dataset is comprised of 35,887 grayscale images, each with a resolution of 48x48 pixels. The dataset is divided into three sets, with 28,709 images allocated for training, 3,589 images for validation, and another 3,589 images for testing purposes. Researchers can access and download this dataset for research purposes.

EmotioNet [53]: EmotioNet is an extensive database that encompasses a collection of one million human faces sourced from the internet. This database is accompanied by annotations for each expression captured in the images. Approximately, 9,50,000 photo

images were annotated through Action Units (AUs) with the help of the intensity of their emotions. The remaining 25,000 images were annotated manually by 11 action units.

AffectNet [54]: The AffectNet database includes more than one million human faces collected from the World Wide Web. Those images are manually labeled for 8 facial expressions (happy, sad, neutral, angry, fear, surprise, disgust, contempt) based on the strength of valence and arousal space. Researchers can access this dataset by email request.

Ferv39k [55]: The Ferv39K is a large-scale video sequence database, which has been collected from real-time video clips from various real-world contexts such as scenes, movies, TV, live shows, and official events. It contains 38,935 video clips of 7 labeled facial expressions.

PEDFE [56]: The Padova Emotional Dataset of Facial Expression (PEDFE) newly created dataset with six universal expressions. It contains 1458 facial images of 56 participants.

UIBVFED-Mask [57]: The UIBVFED-Mask dataset is an extended version of the UIBVFED dataset. Recognising facial expressions from occlusion is a challenging task due to the loss of significant facial expression information. This dataset contains 640 images of 32 facial expressions of 20 participants.

2.5 Facial Expression Recognition Techniques in Machine Learning

The conventional FER approach using ML techniques comprises four key methods: face detection, pre-processing, feature extraction, and emotion classification or recognition, as illustrated in Figure 2.4

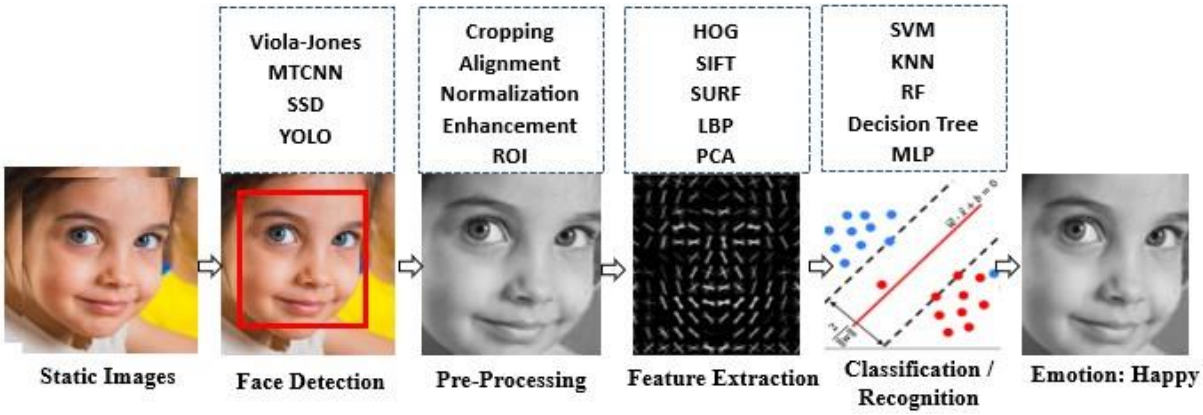

Figure 2.4. Conventional FER in Machine Learning

2.5.1 Face Detection

Face detection holds a significant role as the initial step in many areas including face alignment, face recognition, face verification, and recognizing emotions from face images or video sequences. It serves as a crucial component in the process of emotion recognition. The primary objective of face detection is to ascertain the presence of a face within an image, as this region is vital for emotion detection as well as facial recognition of individuals, as depicted in Figure 2.5. The accurate detection of faces lays the foundation for subsequent steps in the process. In addition, face movement has been tracked in the video sequence. It is part of object detection used in many places like security, biometrics, personal safety, and so on. Face detection and facial recognition are not the same but are interrelated. Face detection allows a system to identify the presence of a human face in an image or video sequence, whereas face recognition can indicate the name of the person in that image.

(a) Original Image (b) Detected Face Image

Figure 2.5. Face Detection from digital images

2.5.1.1 Different Face detection approaches

Different face detection methods, including knowledge-based, feature invariant, template matching, and appearance-based techniques, are employed for detecting faces, as depicted in Figure 2.6. These methods encompass a range of approaches used to accurately identify and locate faces within images or video sequences.

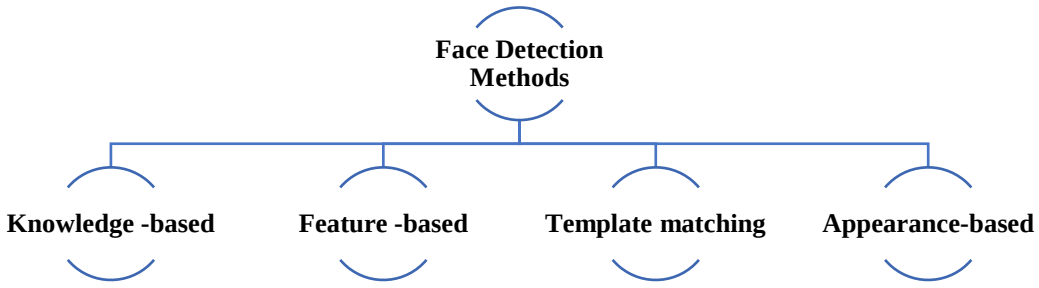

Figure 2.6 Type of Face Detection Approach

A. Knowledge-based

The knowledge-based method follows a set of rules developed by humans. Firstly, facial features such as the nose, eyes, mouth, shape, size, texture, etc., are obtained from images. Secondly, this type of face detection process relies solely on predefined rules, which are performed based on prior knowledge of facial geometry created by human-crafted knowledge. However, these methods prove to be less effective in real-world scenarios, such as faces with illumination, pose variation, and diverse facial features. Nevertheless, this method was found to be useful for front-face images and in well-controlled environments.

B. Feature-based

Feature-based methods extract structural features from images. These structural features include skin color, shape, texture, and facial local features. Other features such as eyes, mouth, eyebrows, and nose are extracted with the help of filters. According to some studies, skin color is considered one of the best features for detecting a face in images. For instance, methods like HOG and Viola-Jones are employed for feature extraction to identify faces or objects in images. However, these features are sometimes corrupted due to factors such as high illumination, face orientation, occlusion, and noise, as they heavily rely on the visibility of specific features Furthermore, these methods may fail to capture the full complexity of facial patterns in diverse datasets. Also, feature boundaries can be weakened for the face, and shades can cause strong edges, collectively making existing feature extraction methods less fruitful in certain scenarios.

C. Template Matching

The template matching method is a straightforward approach that detects faces through the correlation between pre-determined face templates and input images using a predefined or parameterized face template. The edge detection model and filters are employed to construct edges in images. However, this method has some limitations. Accuracy is affected by real-world scenarios with diverse conditions, and computation time increases when searching for faces in larger images. Additionally, it suffers from overlapping faces with complex backgrounds, leading to a higher false-positive rate. Furthermore, selecting an efficient template for various conditions is less robust, as different scenarios may require different templates. Moreover, illumination on the face image also impacts the performance of the template matching method, resulting in non-face region detection.

D. Appearance-based

The appearance-based method identifies relevant patterns and features from facial images using a machine learning algorithm instead of explicit templates. Commonly used techniques in face detection include Haar-like features, Principal component analysis (PCA), SVM, Eigenface, Hidden Markov Model, Naïve Bayes Classifier, and CNN. Furthermore, this method is considered robust compared to other face detection techniques, especially in challenging real-world environments. Nowadays, most face detection techniques are developed based on this approach. However, it often requires a large and diverse set of annotated face and non-face images for training; otherwise, it may suffer from overfitting. Additionally, it may struggle with low-quality image resolution, partial occlusion, noise, blurriness, or compression, and the computational cost is high when developing a face detection model with large real-world images using deep learning techniques.

Table 2.2 Existing Face Detection Techniques from the year 2001 to 2023

Ref	Techniques	Dataset	Accuracy / Result
[58]	You Only Look Once (YOLO), Vgg-16	FDDB, Real-Time Live Video	Achieved 95% average precision
[59]	Dual-Branch Center Face detector (DBCFace)	AFW, PASCAL face, FDDB, WIDER FACE	Achieved 90.34% accuracy
[60]	Soft-NMS, Resnet-50	FDDB	Achieved 94.2% accuracy
[61]	Improved AdaBoost	Real-Time Data	Reduced false detection rate
[62]	MTCNN	WIDER FACE	Obtained 85.7% results
[63]	YOLOv3, Darknet-53	WIDER FACE, FDDB,	Achieved 93.57% accuracy
[64]	Improved MTCNN	MIT, Casia, NICE-II	Achieved 98% accuracy
[65]	You Only Look Once (YOLO)	WIDER FACE, Celeb Faces, FDDB	Achieved efficient face detection time than the traditional algorithm
[66]	Three-category face detector, Fast R-CNN	WIDER FACE, FDDB	Obtained 97% true positive rates.
[67]	Haar Cascade	Instagram Selfie Images	Achieved 71.48% accuracy
[68]	Single-Stage Joint	WIDER FACE	Achieved 56.66% average precision.
[69]	Haar Cascade	Open internet images	Achieved a Positive Prediction Value (PPV) of 98.01 %
[70]	Region-based Fully Convolutional Networks (R-FCN),	WIDER FACE dataset, FDDB dataset	Achieved True Positive Rate 98.99%
[71]	Viola-Jones	MIT, FERET	Achieved 98.97% accuracy
[72]	Compact CNN	FDDB	Achieved high speed compared with traditional GPUs and CPUs

Table 2.2 shows the comprehensive study of the existing face detection methods that have been widely used in recent years. This analysis considers the factors related to the type of techniques used in both pre-processing and face detection, the name of the datasets used for training and testing the model, the platform environment to develop the face detection model, and finally permeance of the developed face detection algorithm. During the last decades, several face detection approaches have been proposed and tested in various environments with many conditions. To begin with, YOLO (You Only Look Once) [58] [63] [65] is a real-time object detection neural network-based algorithm used in face detection. It contains three techniques namely, residual block, bounding box regression, and intersection over union (IOU). For face detection, YOLO is trained with a larger dataset for face detection like FDDB, Wider face, and Celeb face benchmarks. The VGG-16 and DarkNet-53 models have been used to extract features from the images before applying the YOLO algorithm. YOLO algorithm detects the various positions, illumination, and different skin complexations in real-time. However, it suffers from the precise location of small and multiple faces, and different scales of face image on the real-time scenario.

Similarly, Dual-Branch Center Face detector (DBCFace) [59], and ResNet-50 [60] are convolution neural network-based algorithms designed to reduce problems in non-maximum suppression (NMS). It has trained with various databases like AFW, PASCAL face, FDDB, and WIDER FACE. It detects the occulted face with higher accuracy. Adaboost [61] is also known as an adaptive boosting machine learning-based ensemble method which is used to find out the strong features and to identify the face in the YCbCr color model. Multi-task Convolution neural networks [62][64] and Viola-Jones [71] algorithms are used to solve the numerous challenges in face detection methods. MTCNN consists of three networks namely the Proposal Network (P-Net) gives more false-positive predictions, the Refine Network (R-Net) which uses the NMS method to reduce the false positive rate, and the Output Network (O-Net) which gives a more accurate face position with five landmark locations of eyes, nose, and mouth corner [62][64]. Those networks are not connected directly but the result of one network is given to the input of another network.

Furthermore, Viola-Jones [71] is a conventional face detection algorithm widely used to figure out a front face in digital images. This method follows the four main steps namely Haar-like feature, integral images, AdaBoost, and cascade classifier. The Haar-like feature is

used to extract features from the images [69]. These features are transformed into pixel values with the help of integral images. AdaBoost algorithm selects the important features from whole extracted features [61]. A final cascade classifier is used to discard the non-face in an image which speeds up the face detection process. Viola-Jones algorithm gives a more false-positive rate when the face angle is over 45^{o} and above. Single-stage joint face detection [68] is a fully convolution neural network. It contains three components: feature pyramid network gets input face and outputs five scale feature map; context head module calculates multi-task loss from feature map and cascade multi-task loss predict the bounding box from the regular anchor. Region-based Fully Convolutional Networks (R-FCN) serve as an object detection framework [66]. This framework incorporates a ResNet with over 101 layers to extract features from facial images, resulting in enhanced accuracy for face detection in benchmark datasets such as WIDER and FDDB. Moreover, R-FCN has demonstrated superior performance in accurately detecting faces compared to other methods.

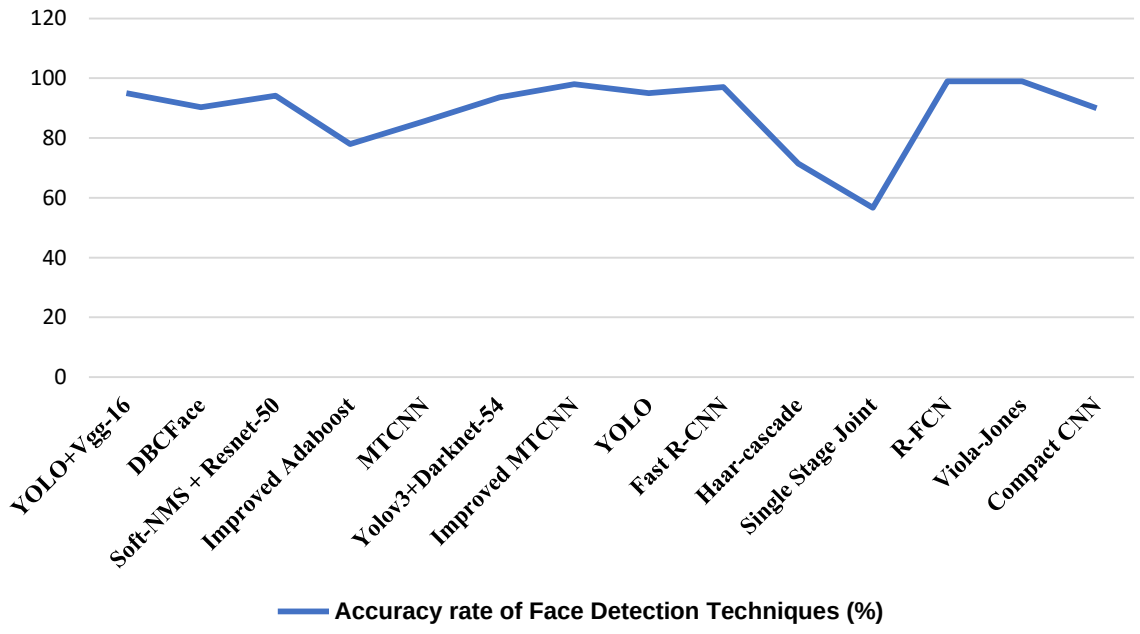

Figure 2.7 Accuracy Rate of Face Detection Techniques

Figure 2.7 illustrates the accuracy rates of various face detection techniques. The x-axis represents the names of the face detection methods, while the y-axis indicates the corresponding accuracy levels obtained from Table 2.2 The graph provides a visual representation of the performance of each face detection technique in terms of accuracy.

2.5.1.2 Challenges in Face Detection

The advancement of computer vision over the last few decades has made research more efficient. The challenges in face detection/face recognition affect the quality of the outcome [73]. The face detection process is difficult due to the various sizes of the face, pose variation, occlusion [74], aging, noise, low resolution, and illumination [75]. Figure 2.8 shows the challenges that occur in face detection and recognition.

Figure 2.8 Typical Challenges in Face Detection

A. Illumination and Pose Variation

To begin with, Illumination refers to lighting conditions and the presence of shadows on the face images that cause a cluttered background for expression images, which can pose challenges in facial detection and recognition. Secondly, the pose of a face can significantly varies due to head movements and changes in viewing angles, leading to inaccuracies or failures in face recognition or detection. Facial expressions constantly change at both macro and micro levels. In 2D facial images, to reduce computational costs, the extraction of spatial features is extremely difficult. Similarly, in 3D images, a side view could affect the system's performance. However, most existing facial expression datasets were collected in a controlled environment, where expressions are static and captured by both professional and non-professional actors. Consequently, the performance of FER is degraded in the real world, where spontaneous and sequential facial images are common.

B. Aging

Similarly, aging involves changes in facial appearance and texture over time, presenting a significant hurdle in accurate detection and recognition due to alterations in features, shapes, lines, and other aspects of the face. To support this claim, studies suggest that when investigating the performance of optical flow and high gradient detection on infants, the developed algorithm showed lower performance on infant facial images than on adults [76]. The reason is that infant skin texture, fatty tissues, and the absence of transient furrows reduce the algorithm's performance. This claim is further emphasized by [77], indicating that different physical appearances, such as skin texture, can affect FER performance. This is also the main reason for not combining multiple facial expression datasets when training the FER model.

C. Partial Occlusion

A partial occlusion occurs when certain parts of the face are blocked, resulting in incomplete input images where the entire face is not available for detection. These factors collectively contribute to the complexities and challenges faced in facial detection and recognition [78]. These challenges extend to natural occurrences such as beards, wearing glasses, hijabs, mustaches, cosmetics, and headscarves. Moreover, in recent times, many studies have been designed to detect faces even when individuals are wearing face masks [79], and this technique could prove fruitful in overcoming these kinds of challenges.

2.5.2 Pre-processing

Pre-processing plays a crucial role in enhancing the efficacy of FER systems before the facial feature extraction process. It involves a series of steps aimed at refining, reducing redundant information, and optimizing the input data features, ultimately improving the accuracy and reliability of the FER system. This process is essential in both conventional machine learning and deep learning methods. Moreover, this phase consists of different types of processes such as face localization, facial landmark detection, normalization, and augmentation. Additionally, it includes various image-enhancing techniques like scaling, contrast adjustment, improving image clarity, histogram equalization, gamma correction, pixel brightness transformation, Fourier transform, and filtering.

A. Face Localization

Face Localization is generally used to detect the region and size of the human face in images or video sequences [80]. This approach removes unrelated background information

that can affect prediction accuracy. Moreover, to detect the face from the input images, Viola-Jones [81], Haar features [82], and the AdaBoost algorithm [83] have been actively used for decades. These algorithms were optimized for speed with integral images. However, Viola-Jones has some shortcomings, including non-robustness in partial occlusion and pose variation. In addition, Region of Interest (ROI) segmentation is one of the most important functions used in face localization to identify and mark the facial organs [84]. Nevertheless, the method fails to implement bounding box regression. This study [85] has explored the issues in bounding box regression using a CNN model to determine if the bounding box accurately fits on the face. The authors applied these steps iteratively until they achieved a significant face location in face images or a sequence of face frames.

B. Face Alignment

On the other hand, face landmark detection (face alignment) is used to mark facial features such as eyebrows, mouth corners, eyes, and lips as shown in Fig. 8. This process is performed after the face detection step. It serves as another pre-processing method for determining the geometrical model of the human face [86][87]. This landmark improves the FER system performance, for instance, the SIFT algorithm is often employed to identify facial features. Subsequently, all facial expressions are aligned using related reference images. The facial landmarks visualize parts of the human face such as the eyes, mouth, nose, eyebrows, and jawlines, as shown in Figure 2.9. Normally, it is used for images or video sequences to detect faces and objects. Also, a comprehensive review is available in this face mark localization for readers [88]. After, the advancement of deep learning methods, face landmark detection became easier and more proved its superiority over existing machine learning-based methods.

Figure 2.9 Facial Landmark Detection

C. Face Normalization

Face normalization is an important pre-processing technique employed to alleviate the impact of irrelevant and redundant information, such as background, hair, and clothing, in order to streamline the detection process [89]. By removing these non-essential elements, face normalization aims to enhance the effectiveness and efficiency of the recognition process, focusing solely on the facial region of interest. In layman's terms, normalization is a method of rotating a non-frontal facial expression to a frontal pose in terms of improving face recognition. For normalization, Euclidean points are being used to measure the position between facial features [90]. Furthermore, several pre-processing methods have been used but ROI and histogram equalization are widely applied in FER pre-processes. Also, cropping and scaling were decided to apply to the face images, with the nose of the facial parts chosen to take as the central axis as well as other points physically involved. Noise reduction is to reduce the noise from facial images as Median Filter (MF), Adaptive Median Filter (AMF), Gaussian Filter (GF), and Bilateral Filter (BF) are often used as filters in FER systems. Gaussian filter is used to resize the image which provides the smoothness of the images. Similarly, Histogram Equalization (EQ) is a pre-processing method used to enhance color contrast in image histograms which separates out the most frequency intensity pixel values [91], which is commonly used in FER.

D. Data Augmentation

Data Augmentation is a technique used to increase the dataset size through computational manipulation, including flipping, cropping, rotation, zooming, scaling, and many more. This approach helps improve the performance of machine learning models, particularly deep learning models. Data augmentation can be implemented in two ways: (1) the *offline approach* is employed when data should be stored in a separate folder after augmentation, and (2) the *online approach* dynamically augments data during training. This approach is widely applied in all deep learning (DL) techniques. However, some existing literature and Facial Expression Recognition (FER) papers have suggested [92] [93] that automatic augmentation can introduce possible biases through a random selection of samples and incorrect augmentation policies. Recently, the AutoAugment approach, using reinforcement learning, has been introduced [94], but it is computationally expensive. Similarly, Population-based Augmentation (PBA) [95] and Population-based Training (PBT) have been presented,

but they have not achieved significant results. In computer vision, robust research on data augmentation is still open to researchers.

2.5.3 Facial Feature Extraction Methods in ML

The feature extraction method of the FER system is the subsequent phase after the pre-processing stage. This stage involves extracting and highlighting useful information from unstructured data. This approach helps reduce potential biases in recognizing facial expressions from a vast number of features, playing a crucial role in Computer Vision. This section comprehensively discusses the existing machine learning feature extraction techniques as follows:

Global and local feature extractors are the two types of feature extractors commonly utilized in digital photographs [96][97]. For image retrieval, object detection, and classification, global descriptors are used. Local descriptors, on the other hand, are utilized in object detection and identification. Moreover, PCA is a dimensionality-reduction approach for extracting local and global level dimensional information [98]. Through multi-channel observation, Independent Component Analysis (ICA) retrieves local features. The feature extraction approach Stepwise Linear Discriminant Analysis (SWLDA) features extracted from forward and backward linear regression. It is determined by the estimated class label F-test values for regression models. The Local Curvelet Transform (LCT) serves as a geometric feature descriptor that effectively captures geometrical features through a wrapping mechanism. This technique extracts features such as mean, median, and standard deviation. Additionally, energy and kurtosis characteristics are obtained by utilizing three-stage directional pyramid representations. These extracted features contribute to a comprehensive understanding of the geometric properties of the analyzed data. The Gabor Filter serves as a texture descriptor utilized for feature extraction, encompassing both magnitude and phase parameters [99]. The magnitude feature provides limited information regarding the arrangement of facial image components, while the phase feature complements it by providing a more comprehensive description. Together, these features offer a comprehensive representation of the texture characteristics present in the face image. LBP is a texture descriptor that is used to retrieve features from images [100]. It generates binary code, which can be obtained by differing the threshold levels between both the center and locality pixel resolution.

The HOG feature descriptor is a window-based technique that leverages gradient filters to extract features from images [101]. Specifically, it focuses on the edge information derived from authorized facial expression images. The HOG descriptor captures visual characteristics, such as smile expressions characterized by curved-shaped eyes. By analyzing the gradients and orientations of local image regions, HOG effectively captures key facial expression features. Similarly, the Active Shape Model (ASM) [102] is a mathematical prototype model that is frequently used to extract feature marks from a facial expression by incorporating local texture characteristics. In order to handle the high-dimensional nature of extracted features, various dimensionality reduction techniques, such as PCA and Linear Discriminant Analysis (LDA), are employed. These techniques aim to reduce the dimensionality of the feature vectors while retaining the most important information. Additionally, different algorithms, such as Viola-Jones and similar ones, are utilized to select the most relevant features, further enhancing the efficiency and effectiveness of the overall facial expression recognition process.

In summary, ML-based feature extraction techniques are not commonly applied in FER systems these days due to the handcrafted feature extraction approach. Additionally, these techniques are not sufficient for extracting subtle, intricate, and complex patterns in facial expressions. Furthermore, the ML approach faces challenges in handling high-dimensional data in the FER dataset, as well as unseen data in real-time scenarios with variations in illumination, pose, and extreme facial expressions. Moreover, it is a time-consuming process to find relevant features to build robust FER systems.

2.5.4 Classification / Recognition Approach of FER in ML

The final stage of the FER system involves emotion recognition or classification, a critical process that assigns specific emotion labels to the input images. This stage is designed to predict emotions such as *happy, neutral, sad, disgust, anger, surprise, fear, boredom, confusion,* and *frustration*, as depicted in Figure 2.10. Typically, FER systems are trained with a core set of six primary emotions—*happy, sad, anger, fear, surprise,* and *disgust*—due to the limited availability of diverse facial expression samples in standard FER datasets.

To achieve a broader emotional understanding, compound facial expressions are incorporated. These are created by combining two basic emotions, resulting in complex

emotional states that go beyond the six primary categories. Overall, FER systems can represent up to twenty-one distinct emotions: the six primary emotions, a neutral expression, twelve compound emotions, and three additional emotions—*awed*, *appalled*, and *hated*. These nuanced classifications enhance the system's ability to interpret a wider range of human emotional experiences.

In addition to classifying static facial expressions, the FER system also considers micro-expressions. Micro-expressions are fleeting, involuntary facial muscle movements that reveal a person's true emotions, often bypassing conscious control. These subtle expressions are typically brief, lasting only about 1/25 to 1/3 of a second, and are difficult to detect without specialized tools. Despite their subtlety, micro-expressions provide valuable insights into concealed or subconscious emotions, making them an essential aspect of understanding human affective states. By capturing both overt and covert expressions, FER systems contribute significantly to applications in psychology, human-computer interaction, security, and emotional intelligence.

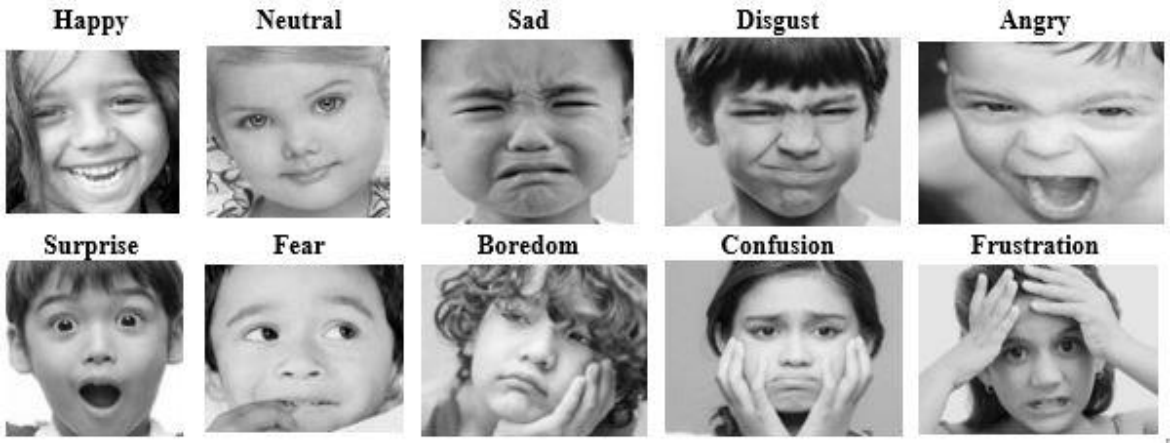

Figure 2.10. Facial Expression in AffectNet [102]

Table 2.3 presents a comprehensive examination of the most commonly applied machine learning algorithms in FER systems. The study takes into account factors such as the year, feature extraction techniques, machine learning algorithms for facial expression classification, types of datasets, the number of expressions, and the accuracy of each method in FER.

Table 2.3. Conventional FER in Machine Learning

Ref	Feature Extraction Techniques	Classifier	Dataset	Emotions Analysed	Accuracy/Result
[103]	Improved Cat Swarm Optimization (ICSO), DCNN,	SVM, Neural network	JAFFE, CK+, PIE, Real-world images	Six emotions	Provided significant accuracy
[104]	Viola-Jones, HOG	SVM	Own Dataset	Four emotions	Speech: 85.72% FER: 92.88%
[105]	Gabor filter	SVM	JAFFE, CK, CK+	Seven emotions	JAFFE:96.30%, CK:94.20%, CK+:94.26%
[106]	Modified Viola-John's	KNN, SVM	JAFFE, LNMIIT, CK+, MMI	Seven emotions	MMI:97.5, JAFFE: 97.65, LNMIIT:99.77, CK+:98.56
[107]	Viola-Jones-face detection, HOG, PCA	SVM, KNN, and MLPNN	CK+	Eight emotions	93% of accuracy
[108]	Viola-Jones,	KNN, SVM, RF, CART	N/A	Six emotions	98.24% of accuracy
[109]	HAAR filters	SVM	CK, CK+	Eight emotions	93.7% of accuracy
[110]	RST-Invariant features and texture features	KNN, SVM, ANN	JAFFE	Six emotions	90% of accuracy
[111]	Viola-Jones, Haar feature, AdaBoost learning, Log-Gabor filters, LBP.	SVM	CK+	Seven emotions	79% of accuracy
[112]	Active Shape Model (ASM) tracker	SVM	CAFÉ	Six emotions	93% of accuracy
[113]	Haar features, Viola, and Jones, AdaBoost Classier (EmguCV, OpenCV)	SVM	Real-time face Stimuli	Four emotions	87.9 % of accuracy

Feature extraction methods play a crucial role in FER systems by isolating and quantifying relevant facial features from pre-processed images. Techniques such as HOG [104], Gabor filters [105], PCA [107], HAAR filters [109], Log-Gabor filters [111], LBP [111], and ASM are commonly utilized for this purpose. These methods help in capturing specific patterns, textures, and geometric structures essential for emotion classification. The application of these techniques was elaborated in the previous section. FER systems aim to identify and classify more than five primary emotions (e.g., happy, sad, anger, fear, surprise, and sometimes neutral or contempt). To train and evaluate these systems, widely recognized datasets such as CK, CK+, JAFFE, MMI, and PIE are used. These datasets provide diverse sets of labeled facial expressions that ensure robustness and generalizability of the models. To detect faces within digital images, traditional methods like the Viola-Jones algorithm leverage HAAR features and the AdaBoost algorithm. These methods are effective in isolating the face region, a critical step, as FER systems focus only on the facial area to process and classify emotions accurately. Identifying the face ensures the exclusion of extraneous background information that could interfere with emotion recognition.

Before the advent of deep learning approaches, ensemble techniques were frequently employed to enhance the efficiency and accuracy of facial emotion recognition systems. By combining multiple feature extraction or classification methods, ensemble techniques leverage the strengths of each approach while mitigating individual weaknesses. This strategy significantly improved performance in scenarios with diverse or challenging datasets, serving as a bridge between conventional machine learning and the more recent adoption of deep learning methods.

Upon closer examination, the Support Vector Machine (SVM) has proven to be a robust supervised learning algorithm widely used for classification tasks in FER [103][104][105]. SVM operates by constructing an optimal decision boundary, referred to as a hyperplane, that separates data points belonging to different classes. Through an iterative process, it fine-tunes this hyperplane to minimize classification errors while maximizing the margin between classes. This approach enables SVM to excel in handling complex classification and regression problems, offering high accuracy in predicting emotions. By leveraging the concept of the Maximum Marginal Hyperplane (MMH), SVM effectively classifies features into six distinct emotions. Similarly, the K-Nearest Neighbor (KNN)

algorithm is another supervised machine learning technique employed in FER. While KNN can be used for both classification and regression, it is predominantly applied for classification. KNN operates by storing all available training data and categorizing new data points based on similarity metrics, such as distance calculations. This method allows for better clarity and faster computations during classification. The training samples are organized in an n-dimensional feature space, enabling efficient analysis and classification of facial expressions. Both SVM and KNN remain foundational algorithms in FER for their simplicity and effectiveness.

Random Forest (RF) is an ensemble learning technique that combines multiple classifiers to address complex problems and enhance model performance. It operates in two phases. In the first phase, the random forest is created by combining multiple decision trees. In the second phase, predictions are made by each individual tree generated in the first phase. RF offers several advantages, including reduced training time compared to other algorithms, high prediction accuracy, and efficient performance with larger datasets. Similarly, one popular decision tree algorithm used within RF is CART (Classification and Regression Trees) [108]. CART employs a tree-based structure and utilizes the if-then-else rule to predict outcomes for data points. This algorithm finds applications in both facial emotion recognition and facial expression recognition tasks. By leveraging the inherent decision-making capabilities of decision trees, CART contributes to accurate predictions and the effective recognition of facial emotions and expressions. Moreover, a nature-inspired algorithm was also applied to optimize the feature search techniques. For instance, Cat Swarm Optimization (CSO) is an intelligent optimization technique inspired by the behavior of cats and operates in two distinct modes: seeking mode and tracing mode [103]. In the seeking mode, the cat assumes a relaxed position, while in the tracing mode, it mimics the behavior of a cat searching for prey. This swarm-based approach draws upon the instincts and behaviors of cats to efficiently explore and exploit search spaces, leading to effective optimization results. CSO is used in the FER system to select the best features to improve FER recognition accuracy. Additionally, particle swarm optimization and Ant Colony Optimization (ACO) [103] have been used in facial emotion recognition systems in the past.

Furthermore, neural network-based techniques have begun to be applied in the FER system. This architecture, consisting of three layers, includes the input layer, which receives

the input data; the hidden layer, responsible for processing and transforming the input; and the output layer, which produces the classification results. This neural network architecture enables the model to effectively learn and extract meaningful features from facial expressions, facilitating the accurate classification of various emotions in this network, parameters such as the classification feature vectors, the dimension of the feature vector, and the total number of classes, such as happy, sad, surprise, neutral, angry, and fear, are used. One of the major strengths of feedforward artificial neural network techniques is the multiple-layer perceptron neural network (MLPNN) [107]. The neurons in the MLP are trained with a backpropagation learning algorithm for classification, recognition, approximation, and prediction.

In summary, ML classifiers face challenges in generalizing to diverse FER datasets and real-world scenarios. Variations in illumination, pose, and facial expressions often lead to poor performance when tested on unseen data. These classifiers typically handle static images independently, making them inadequate for analyzing temporal facial image sequences in videos. A significant limitation is their reliance on large, well-annotated FER datasets, which are challenging to collect. Class imbalance in datasets can further reduce accuracy, as underrepresented emotions are not effectively learned. Additionally, ML models struggle with recognizing emotions across diverse individuals, as variations in age, gender, and ethnicity affect facial expressions. These issues highlight their difficulty in achieving consistent performance. As a result, their applicability to dynamic, real-world environments remains limited. Overcoming these barriers requires more robust techniques that account for temporal and demographic variations.

2.6 Facial Expression Recognition Using Deep Learning

Deep learning is a field within machine learning that draws inspiration from the functioning of the human brain, specifically neural networks. It aims to develop algorithms and models that can learn and make predictions by mimicking the complex processes of the brain. There are several kinds of deep learning techniques such as Artificial Neural Networks (ANN), Autoencoders, RNN, and Reinforcement learning are shown in Figure 2.11. Among all, CNN or ConvNets are frequently adapted in FER image processing, The main advantage of this method is combining the feature extraction and classification parts, which greatly reduces the handcrafted feature extraction process and its challenges. The following

subsections present an overview of CNN and Transfer Learning (TL) techniques that are commonly employed in FER systems. In addition, state-of-the-art techniques like autoencoder, GAN, Bi-LSTM, RNN, reinforcement learning, and ensemble methods have recently been applied in the FER system. These techniques increase the accuracy of emotion recognition when compared to conventional approaches and are also highly efficient for extracting spatiotemporal features in a sequence of facial expressions.

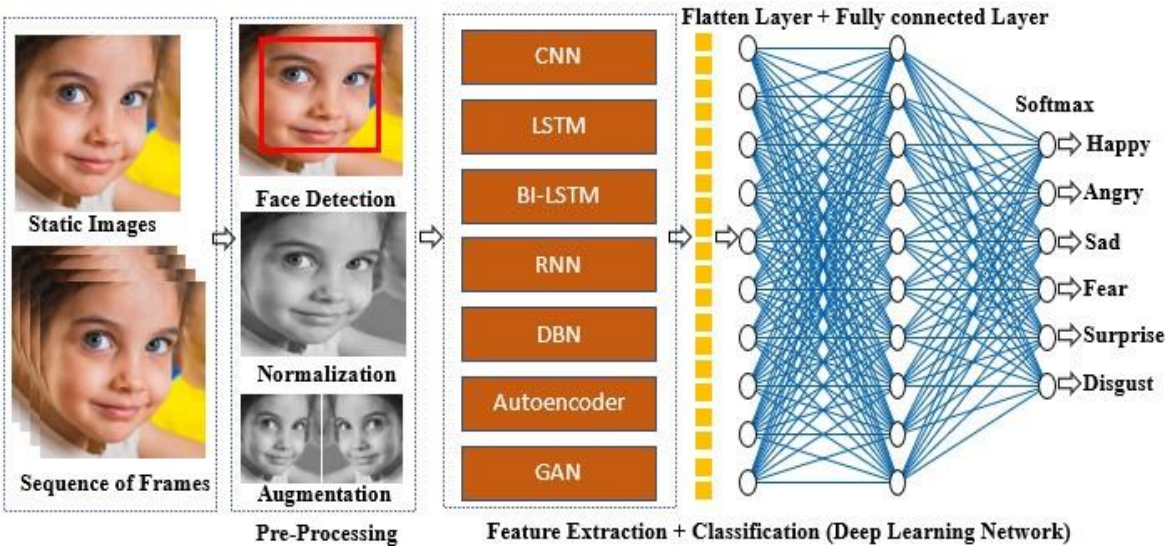

Figure 2.11 Deep Learning-based FER

Figure 2.11 outlines a deep learning-based FER system. It starts with input data (static images or sequences of frames), followed by pre-processing steps like face detection, normalization, and data augmentation to ensure consistent input. In the feature extraction stage, models such as CNN, LSTM, Bi-LSTM, RNN, DBN, Autoencoders, and GANs extract spatial and temporal features. The extracted features are then passed through a flattened layer, which converts them into a one-dimensional vector and subsequently processed by fully connected layers for classification. Finally, the Softmax layer predicts the probability of the image or sequence belonging to one of the predefined emotion categories, such as happy, angry, sad, fear, surprise, or disgust. This pipeline highlights a sophisticated approach that leverages spatiotemporal features and cutting-edge deep learning models to improve FER performance in dynamic environments.

Table 2.4 Deep Learning-based FER Approach

Ref	Techniques	Dataset	Emotions	Accuracy/Result (%)
[114]	DCNN (VGG-16)	KDEF, JAFFE	Afraid, Angry, Disgust, Happy, Neutral, Sad, Surprised	KDEF: 93.47 JAFFE: 100
[115]	CNN, RNN, ConvLSTM	FER-2013	Angry, Disgust, Fear, Happy, Sad, Surprise, Neutral	CNN: 65 RNN:41
[116]	CNN, OpenCV	FER-2013	Neutral, Happy, Sad, Angry, Surprised, Disgusted	93.95
[117]	DCNN	Real-time dataset	Angry, Happy, Neutral, Sad, Surprise	78.04
[118]	CNN	FER-2013	Angry, Disgust, Fear, Happy, Sad, Surprise, Neutral	Training:97 Testing:57.4
[119]	CNN	FER-2013	Angry, Disgust, Fear, Happy, Sad, Surprise, Neutral	62%
[120]	CNN, Haar-Cascade Classifier	FER-2013	Angry, Disgust, Fear, Happy, Sad, Surprise, Neutral	Experimental done with different epochs.
[121]	CNN	FER-2013	Angry, Disgust, Fear, Happy, Sad, Surprise, Neutral	Accuracy 79.8
[122]	CNN	FERC-2013, JAFFE	Angry, Disgust, Fear, Happy, Sad, Surprise, Neutral	FERC:70.14 JAFFE: 98.65

Ref	Techniques	Dataset	Emotions	Accuracy/Result (%)
[123]	CNN, Viola-Jones, Haar feature	NIMH-ChEF, CAFÉ, AM-FED, and EmoReact	Neutral, Happy, Sad, Surprise, Fear, Disgust, and Anger	46.05
[124]	CNN	FER-2013	Angry, Disgust, Fear, Happy, Sad, Surprise, Neutral	70
[125]	DCNN, OpenCV	ADFES-BIV, WSEFEP	Happy, Sad, Anger, Surprise, Disgust, Fear, Neutral, Pride, Contempt, Embarrassment	95.12
[126]	CNN, DNNs	Karolinska Directed Emotional (KDEF)	Afraid, Angry, Disgusted, Happy, Neutral, Sad, Surprised	86.73 5
[127]	DNNs	CK+, JAFFE	Sad, Happy, Surprised, Angry, Neutral, Disgust, Fear	JAFFE: 95.23 CK+: 93.24
[128]	CNN	CK+, KDEF	Anger, Disgust, Fear, Happy, Sa, Surprise, Neutral	Testing: 97.53 JAFFE:97.53
[129]	CNN	FER-2013	Angry, Disgust, Fear, Happy, Sad, Surprise, Neutral	57.1
[130]	CNN	FERC-2013, CK+	Angry, Disgust, Fear, Happy, Sad, Surprise, Neutral	Around 90+ accuracy

State-art-of-the Techniques

Ref	Techniques	Dataset	Emotions	Accuracy/Result (%)
[131]	Feature separation model exchange-GAN	Multi-PIE, FACES, Oulu-CASIA	Sad, Happy, Surprised, Angry, Neutral, Disgust, Fear	Multi-PIE: 91.08 FACES: 95.24 Oulu-CASIA:86.33
[132]	Residual Variational Autoencoder	Affectnet	Neutral, Happy, Sad, Surprise, Fear, Disgust, and Anger	98.0 %
[133]	CNN-LSTM	FER-2013, CK+	Angry, Disgust, Fear, Happy, Sad, Surprise, Neutral	FER-2013:78.2 CK+: 99.7
[134]	GAN	Multi-PIE, MMI, RAF-DB	Sad, Happy, Surprised, Angry, Neutral, Disgust, Fear	Multi-PIE: 93.66 MMI: 76.44 RAF-DB: 89.01
[135]	Ensemble Classifier	JAFFE, TFEID, Moroccan, Caucasian	Sad, Happy, Surprised, Angry, Neutral, Fear	JAFFE: 86.67 TFEID: 83.19, Moroccan: 89.47, Caucasian: 86.36
[136]	Multi-scale convolutional and residual block-based DCNN	FER2013, JAFFE, CK+, KDEF, RAFDB	Neutral, Sad, Happy, Surprised, Angry, Neutral, Fear, Disgust	FER2013: 80 JAFFE:99 CK+:98 KDEF:97, RAFDB:87

[137]	Frequency neural network (FreNet)	CK+ OULU KDEF	Angry, Disgust, Fear, Happy, Sad, Surprise, Neutral	CK+: 98.91 OULU: 88.33 KDEF: 91.22
[138]	Mobile-Net	FER+ RAF-DB	Angry, Disgust, Fear, Happy, Sad, Surprise, Neutral	FER+: 88.11 RAF-DB: 84.49
[139]	Bi-LSTM	SAVEE, RAVDESS, and RML	Angry, Disgust, Fear, Happy, Sad, Surprise, Neutral, Calm	SAVEE: 99.75 RAVDESS: 94.99 RML: 99.23
[140]	Convolutional 3D	CK+, MMI Oulu-CASIA	Angry, Disgust, Fear, Happy, Sad, Surprise, Neutral	CK+: 93.45 MMI: 84.53 FERA: 93.45
[39]	Temporal Relational Network (*TRN*), Multi-Layer Perceptron	DISFA+	Angry, Disgust, Fear, Happy, Sad, Surprise, Neutral	Training accuracy: 92.7 89.4 and 86.6
[41]	MobileNet-v1 EfficientNet-B0 EfficientNet-B2	AffectNet	Sad, Happy, Surprised, Angry, Neutral, Disgust, Fear	Positive:73.2 Negative:70.7 Neutral: 64.3
[42]	Hybrid- CNN	DAiSEE CK+ JAFFE	Sad, Happy, Surprised, Angry, Neutral, Disgust, Fear	DAiSEE: 53.4 CK+: 71.4 JAFFE: 99.95
[43]	Heuristic multimodal real-time emotion recognition approach	CK+, BU-3DFE	Facial expression analysis, vocal intonations, and gesture	Facial detection: 84.25 Hand gesture: 92.70 Voice recognition: 82.26

Table 2.4 presents a comprehensive study of a frequently used deep learning algorithm in FER systems. This study considers factors related to deep learning techniques, datasets, the number of emotions, accuracy, and results across various datasets. The conventional FER system typically involves three main methods: face detection, feature extraction, and the classification of emotions such as happy, fear, sad, anger, neutral, surprise, and fear. First, pre-processing techniques are applied to facial images, including histogram equalization and conversion to grayscale. Then, face detection techniques such as Viola-Jones, Haar cascade, and multi-task convolutional neural networks are employed to detect a face in images or video frames. Finally, Convolutional Neural Networks extract useful information and classify emotions based on labeled datasets.

Based on analysis, CNN [118][119][121][124][128] is a subfield of neural networks, comprised of two major blocks: feature extraction and classification. The CNN architecture includes layers such as the convolution layer, pooling layers, dropout, activation function, and fully connected layers, along with batch normalization and regularization, which are employed when CNN encounters an overfitting problem [116][125][127]. The convolution layer, the lowest layer, is used to extract various information from the input images using an (MxM) filter, and the result is known as a feature map. Feature maps provide information about the edges and corners of an image. The pooling layer's primary purpose is to reduce the dimensionality of the feature map and computation expense. The fully connected layer, responsible for connecting multiple layers, consists of bias, weight, and neurons. Typically, it is placed before the output layer of the CNN. The dropout layer is used to address the overfitting problem in the CNN architecture. The activation function initiates the connection between layers using neurons. Several commonly used activation functions include Softmax, ReLu, TanH, and Sigmoid. To build the CNN model, the dataset can be divided into three sets: training, validation, and testing, along with batch size and epochs [129][130].

Typically, training a neural network requires a massive amount of data to achieve significant accuracy. In deep learning, data augmentation techniques, such as flipping, rotating, scaling, cropping, translation, and adding Gaussian noise, are used to increase the size of the dataset when the DL model suffers from overfitting issues. TensorFlow, Keras, PyTorch, and various Python libraries are commonly employed to develop the CNN architecture. FER utilizes pre-trained Deep Convolutional Neural Network (DCNN) models

through appropriate transfer learning, such as VGG-16 [114], ResNet, DenseNet, and Inception, which are trained with large datasets (e.g., ImageNet) containing different classes. CNN yields better results when combined with transfer learning as the base model [114]. Also, fine-tuning is a crucial step in transfer learning-based FER systems, and carefully chosen techniques are employed to fine-tune the proposed model for better results. Moreover, optimizers such as Adam, AdaGrad, and RMSProp, along with the learning rate, are used to select relevant features from the available ones in this optimization process. Furthermore, in recent years, state-of-the-art techniques such as Bi-LSTM [139], Autoencoder [132], and GAN [134] have been applied in FER to enhance accuracy and overcome challenges, such as the vanishing gradient problem in the conventional CNN approach [137]. Also, this approach is efficient for the extraction of spatiotemporal features from video sequences.

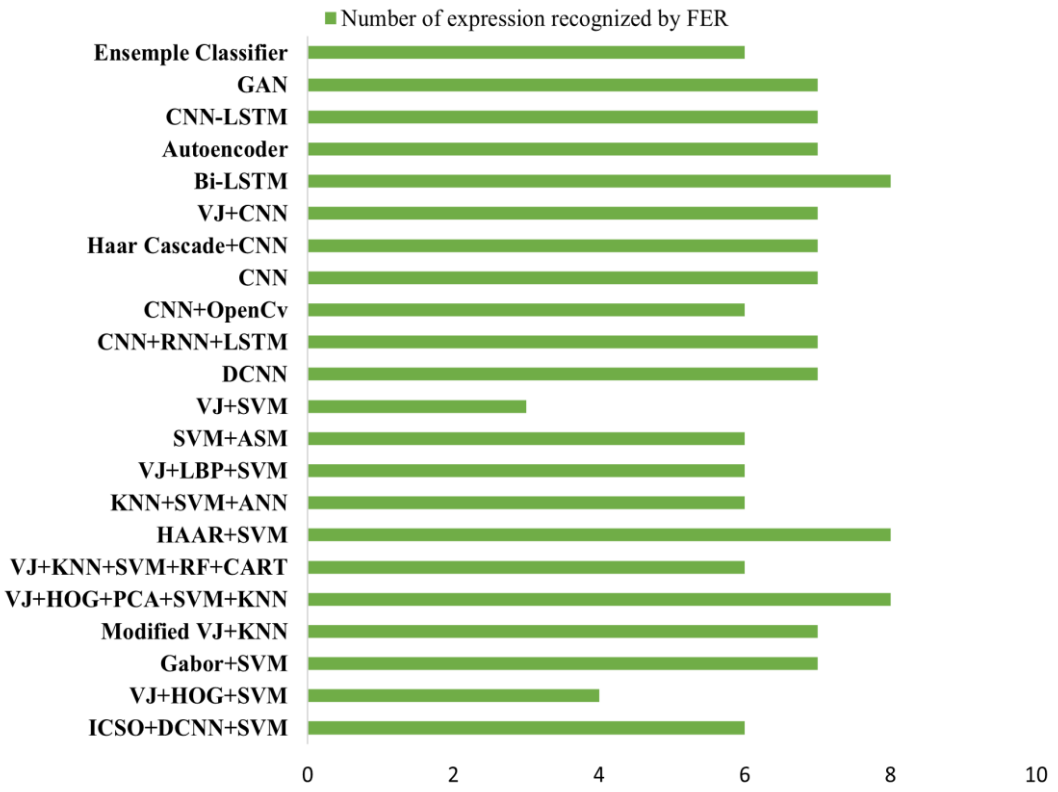

Figure 2.12 Number of Facial Expressions commonly used in FER systems

The number of facial emotions recognized by different machine learning and deep learning techniques is captured and summarized in Figure 2.12. On the x-axis, the number of expressions recognized by FER methods is represented, while the y-axis displays the names of FER methods within the machine and deep learning domains. Based on the analysis

conducted in this survey, it is observed that the majority of papers recognize up to eight facial expressions as the maximum number of emotions in their FER systems. This comparison provides insights into the range and diversity of emotions recognized by different FER methods various machine learning and deep learning approaches.

2.7 Explainable AI Techniques

Explainable AI (XAI) refers to a field within AI research that focuses on developing machine learning models with transparent and interpretable mechanisms [146]. The primary objective of XAI is to ensure that these models can be easily understood and justified by humans, thereby enabling individuals to comprehend the rationale behind the model's predictions or decisions. This is particularly crucial in domains like medical diagnosis or criminal justice, where the ability to elucidate the reasoning behind AI-driven outcomes is essential for trust, accountability, and ethical considerations. Many contemporary machine learning models, such as deep neural networks, are often regarded as "black boxes" due to their complex internal workings, making it challenging for humans to interpret their outputs. XAI strives to address this challenge by enhancing model transparency, interpretability, and accountability, fostering greater confidence and reliability in AI applications across diverse sectors. The following are some important techniques commonly used in ML and DM models.

- **Feature Importance:** This technique involves analyzing the importance of different features or inputs in the model's decision-making process. Methods like permutation importance, SHAP (SHapley Additive exPlanations), and LIME (Local Interpretable Model-agnostic Explanations) are commonly used for feature importance analysis.
- **Interpretable Models:** Using inherently interpretable models such as decision trees, linear models, or rule-based systems instead of complex models like deep neural networks. While these models may sacrifice some predictive power, they offer more straightforward explanations for their predictions.
- **Saliency Maps:** In the context of image data, saliency maps highlight the most influential regions of an input image that contribute to the model's prediction. Techniques like Grad-CAM (Gradient-weighted Class Activation Mapping) and CAM (Class Activation Mapping) are examples of saliency map generation methods.

- **Attention Mechanisms:** Commonly used in sequence-to-sequence models like transformers, attention mechanisms provide insights into which parts of the input sequence are most relevant at each step of the model's processing, aiding in understanding the model's decision process.
- **Counterfactual Explanations:** Generating counterfactual instances involves creating modified versions of input data to understand how changes in input features would affect the model's predictions. This technique helps in identifying factors that influence the model's decisions.
- **Model Distillation:** Simplifying complex models by training smaller, more interpretable models to mimic the behavior of larger models. This approach retains predictive accuracy while improving interpretability.
- **Explanation Generation:** Automatically generating human-readable explanations for model predictions using techniques like natural language generation (NLG). These explanations help users understand the reasoning behind the model's outputs.
- **Model-Specific Techniques:** Some models, such as decision trees or Bayesian networks, have built-in interpretability features that allow for a direct understanding of how inputs lead to outputs.

These XAI techniques play a crucial role in ensuring that AI systems are not just accurate but also understandable and explainable, promoting trust, accountability, and ethical use of AI technologies.

Manresa-Yee et al. [142] utilized Explainable AI (XAI) techniques to investigate how gender influences learning and evaluated which facial expressions exhibit similar patterns across face regions crucial for classification. Their findings revealed common regions in certain expressions for both genders, such as happiness, albeit with varying intensities. Conversely, expressions like disgust showcased significant differences in essential face regions between genders. These insights are pivotal for enhancing FER systems and comprehending potential sources of inequality in this domain.

Ramis Guarinos et al. [145] conducted a facial expression recognition study targeting individuals with intellectual disabilities, aiming to integrate this technology into a social robot context. They trained two prominent neural networks using five facial expression databases and evaluated their performance using two additional databases comprising

individuals with and without intellectual disabilities. The authors employed Explainable AI (XAI) techniques, specifically LIME and the RISE (Randomized Input Sampling for Explanation) approach, to provide explanations for the model's prediction results.

Del Castillo Torres et al. [141] study presented the results of comparing LIME and CEM applied over complex images such as facial expression images. LIME highlights the areas of the image that contribute to a classification. As an alternative to LIME, the CEM (Contrastive Explanation Method) method focuses on providing explanations for the predictions of machine learning models by contrasting them with alternative outcomes or scenarios. The key idea behind CEM is to generate explanations that highlight why a particular prediction was made by the model, as opposed to alternative predictions that could have been made.

Deramgozin et al. [143] introduced a hybrid AI explainable framework (HEF) consisting of a primary functional pipeline that includes a CNN for classifying input images. They also incorporated an explainable pipeline utilizing Facial Action Units and application-agnostic models like LIME. This combination of pipelines provides more comprehensive data to explain the obtained results and bolster the decisions made by the main functional pipeline. The authors validated HEF using the CK+ dataset, demonstrating highly promising results in terms of the explainability of the outcomes achieved.

Kandeel et al. [144] introduced a CNN model for implementing a FER approach to detect emotions in drivers. However, they acknowledged that like many other deep learning approaches, their model lacked transparency and interpretability. To overcome this limitation, the authors employed Explainable AI techniques that generate interpretations for model decisions, providing human-understandable representations. They utilized two XAI visualization methods to support their decision on the architecture of the proposed FER model. This model achieved impressive accuracies of 92.85%, 99.28%, 88.88%, and 100% for the JAFFE, CK+, KDEF, and KMU-FED datasets, respectively. These results highlight the effectiveness of incorporating XAI techniques in enhancing the interpretability and performance of FER systems.

2.8 Research Gap Identified in Systematic Review

From the detailed study above, it is evident that some researchers have proposed using FER in learning environments to assess learners' engagement and provide appropriate

feedback. Similarly, conventional learning content, such as multimedia-based material, is highly effective, though it does not offer an immersive learning experience for students. An essential research gap has been identified.

- Conventional FER systems are designed using CNNs, RNNs, and LSTM networks. CNNs are utilized to extract spatial features from each frame, which are then fed into RNNs or LSTMs to capture the temporal correlations between frames in the input video. However, these systems often struggle to effectively model and encode the spatiotemporal relationships among the input video frames.
- Limited focus on real-time FER systems that can accurately and efficiently detect learners' facial expressions during live interactions in learning environments.
- The supervised approach in FER often demands substantial amounts of labeled data for training, which might not be readily available in FER datasets. This scarcity of labeled data can hinder the training process. Additionally, existing FER benchmark image datasets used in CNNs can lead to overfitting issues due to an inadequate representation of each facial expression category, impacting the model's generalization ability.
- Conventional learning content primarily engages two senses—vision and audio—occasionally incorporating haptic effects, while often neglecting other senses like olfactory and gustatory. This limitation hinders the creation of a fully immersive learning experience, which could significantly enhance engagement and effectiveness. Moreover, innovative approaches are needed to analyze learners' engagement and dynamically adjust the learning content based on their emotional states.

2.9 Chapter Summary

This chapter discusses the background studies on existing FER approaches that have shown significant results, although they have not been effectively analyzed in real-time environments. Additionally, there is a lack of publicly available data on learning-related facial expressions. The conventional FER approach encompasses face detection, pre-processing methods, feature extraction techniques, and emotion classification using ML classifiers. However, these hand-crafted feature extraction techniques prove inefficient in handling complex facial features, and classifiers often struggle to predict accurate emotions in real-time scenarios. On the other hand, DL-based techniques in FER have demonstrated

significant improvements in the extraction of relevant features for facial expression recognition, eliminating the need for manual intervention in the feature extraction process. Consequently, DL methods show robustness in handling complex facial images, including challenges such as illumination, pose variation, and extreme facial expressions. However, it's worth noting that these methods consume more time in both training and testing, necessitating the use of GPU and TPU processors. Moreover, ensemble DL techniques, as well as the use of GANs, LSTM, and Autoencoders, have shown promising accuracy in current FER systems based on this analysis. Also presented available approaches of XAI techniques to validate the performance of DL models.