Journal of Artificial Intelligence and Data Science (JAIDA)
e-ISSN: 2791-8335
https://dergipark.org.tr/pub/jaida

Vol. 5, No. 2, pp. 75-88, 2025
Research Article
Received: 23/07/2025 Accepted: 18/11/2025

Automatic Classification of Basic Emotions Using Deep Learning Techniques

Özen ÖZER [1,*], Nadir SUBAŞI [2]

Abstract

This study presents a comparative analysis of deep learning architectures for the automatic classification of seven basic emotions (anger, disgust, fear, happiness, sadness, surprise, and neutrality) from facial expressions. A key investigation point is the impact of data representation format on model performance. To this end, the FER-2013 dataset was utilized in two distinct formats: a directory structure containing 28,709 images and a CSV file comprising 35,887 images represented as pixel values. For each data format, we developed and evaluated two types of models: a standard Convolutional Neural Network (CNN) and a hybrid CNN-Long Short-Term Memory (CNN-LSTM) architecture. The hybrid model was designed to capture potential temporal dynamics by processing features extracted by the CNN as sequential data. Experimental results demonstrated that the choice of both data format and model architecture significantly influences performance. The CNN-LSTM model, particularly when trained on the CSV data, achieved superior accuracy, highlighting its effectiveness in learning complex feature sequences. This comprehensive comparison provides valuable insights into optimal data preparation and model selection for robust emotion recognition systems, with applications in healthcare, education, and security. Ethical considerations regarding data privacy and algorithmic bias were integral to the study

Keywords: Artificial intelligence; deep learning; diverse data set; emotion classification; facial expression analysis; long – short term memory.

1. Introduction

Emotion recognition has attracted immense interest in recent times in psychology, human-computer interaction, and AI disciplines due to its tremendous potential to revolutionize both the way humans interact with machines and interpret the emotions of real-world applications. Current systems have several limitations in generalizability and adaptability to natural conditions.

Deep learning has emerged as a promising solution for automatic emotion detection and has gained widespread attention, especially with techniques like convolutional neural networks (CNN) and recurrent neural networks (RNN). For developing temporal patterns in facial expressions, this work has employed Long Short-Term Memory (LSTM) networks, which are a specialized form of RNNs. It would be an appropriate model for tasks in which data is to be analyzed sequentially. By combining LSTMs with innovative preprocessing techniques such as data augmentation and transfer learning, this system achieves state-of-the-art performance.

The core of the system is to create an automatic system that can correctly identify and classify fundamental emotions from one scenario to another. Major objectives of this research work can be summarized as follows:

• Developing a robust LSTM-based architecture that can handle dynamic emotional expressions.

• Reliable performance at all light spectrum conditions, demographical groups, and all facial orientation positions.

• Real-time operation capabilities for seamless integration into applications such as surveillance, healthcare, and education.

• Ethical considerations, most notably user consent, data protection, and the discourse on potential biases.

Over the past couple of years, emotion recognition has grown into one of the key research areas in AI and computer vision. This literature review is meant to investigate existing studies, recent developments, and gaps in the research regarding image-based emotion recognition.

Emotion recognition studies base their foundation on the basic theories of emotion described in psychology.

*Corresponding author

***Özen ÖZER**; Department of Mathematics, Faculty of Science and Arts, Kırklareli University, Kırklareli, Türkiye; e-mail: ozenozer39@gmail.com;

0000-0001-6476-0664

Nadir SUBAŞI; Department of Computer Programming,Kırklareli University, Kırklareli, Türkiye; e-mail: nadir.subasi@gmail.com ;

0000-0002-5657-9002

The pioneering work by Ekman identified six basic emotions: happiness, sadness, anger, fear, surprise, and disgust [1, 2].These are commonly taken as a basis in the formulation of many emotion recognition systems. Later, Plutchik introduced the wheel of emotions in 1980, arguing for a more complex spectrum of emotions [3].

More recently, other dimensional models have been used in emotion recognition studies, such as Russell's 1980 valence–arousal model [4].While most of the earlier works on emotion recognition have used handcrafted features with traditional machine learning algorithms, a classic example would be in [5], which combined LBP with SVM for classifying facial expressions. Similarly, Bartlett et al. [6] proposed systems for emotion recognition based on Gabor filters and the AdaBoost algorithm. While such approaches showed promising results in the controlled environment, their performance degraded when taken into the real world.

Deep learning has brought a paradigm shift in emotion recognition. Tang et al. [7] used Deep Belief Networks to classify facial expression, which led to remarkable improvements. Specifically, CNNs have grown out to be highly effective in emotion recognition. For instance, using CNN, Mollahosseini et al. [8] reported high accuracy on a number of datasets.

Recently, transfer-learning approaches have also been favored. He et al. [9] reported very promising performance for the emotion-classification task using the pre-trained AlexNet model. Lightweight architectures such as EfficientNet have also been applied for emotion recognition with both good accuracy and computational efficiency [10]. However, a serious bottleneck in emotion-recognition research is the lack of large-scale and diverse databases under realistic conditions. Another challenge lies in issues such as cross-cultural differences in emotional expression, micro-expressions, and context. Tan and Le [10] give an elaborate review of these problems along with presenting future directions in research.

Real-time emotion recognition has recently become a hot spot in research, especially in mobile and embedded systems. Zhang et al. [11] proposed a low-latency emotion-recognition system based on a MobileNet architecture, while Arriaga et al. [12] designed real-time emotion-recognition applications that could run on mobile devices thanks to deep-learning models (e.g., CK+: Lucey et al. [13], FER-2013: Goodfellow et al. [14], AffectNet: Mollahosseini et al. [8]).

The emerging prevalence of emotion-recognition technologies has increasingly made their application an ethical and privacy concern. Stark and Hoey [15] conducted an in-depth study regarding the use of ethical emotion-recognition systems and scenarios of misuse.Development of a Code of Ethics on Data Collection, Usage, and Storage remains an important research area.

Over recent years, there has been significant progress in image-based emotion recognition. Deep-learning techniques, real-time applications, and large-scale datasets have become focal points of research.

Some interesting research directions in this area include the following;

Cao et al. [16] presented a practical transfer-learning algorithm for facial verification and improved model adaptability across domains and tasks. Chechik et al. [17] explored large-scale online learning for image-similarity ranking. Chen et al. [18] discussed anomaly detection using relational analysis of image data. Chopra et al. [19] proposed a unique similarity metric for face verification. These references represent the wide range of insights into computer vision and machine learning from recognizing difficulties regarding emotion recognition by Dhall et al. [20] and cross-pose facial recognition by Ekenel et al. [21] to state-of-the-art image-processing techniques by Gonzalez et al. [22] and the latest deep-learning approaches by He et al. [9] and Wang et al. [23].

The chapter entitled "Enhancing Law Enforcement through Pose-Based Facial Recognition and Image Normalization Techniques" by Özer [24] contributes to research on new methods for face-recognition enhancement in criminal investigations. This work discusses the effect of improving face recognition and retrieval by similarity tasks [25], effectively minimizing intra-class variance while maximizing inter-class distance. An investigation on emotion classification based on deep learning with small training datasets is conducted using transfer learning algorithms [26]. A new discriminative locality-alignment-based unsupervised feature-selection method is introduced [27]. Densely connected convolutional networks, known as DenseNet, were developed to enhance the flow of information and propagation of gradients [28].

The book provides foundational knowledge in digital image processing [29], and a novel fast exact search method in Hamming space is introduced [30]. The paper presents an information-theoretic metric-learning algorithm [31]. Multi-view face-recognition techniques are discussed [32]. The book provides an in-depth discussion of multi-view geometry in computer vision [33]. The paper provides a thorough look at diffusion models in artificial intelligence [34]. The introduction of LSTM networks changed the game for processing sequential data [35]. An online metric-learning algorithm is developed [36]. A comprehensive review of neural-network-based face-recognition techniques is presented [37]. A kernelized locality-sensitive hashing method is created [38]. The exploration of metric-learning techniques focuses on improving feature representation [39]. A

review of deep facial expression recognition highlights recent advancements [40]. Diffusion models are extensively reviewed [41].

A deep-learning-based fusion face-recognition approach is designed [42]. A multi-metric learning algorithm is proposed [43]. An image analysis technique based on angles is proposed [44]. A fuzzy-rough-image-based cluster optimization approach is proposed [45]. The study dives into deep neural networks for recognizing facial expressions [46]. A fresh approach to personalized image search is introduced [47]. A deep-learning-based face-recognition framework is introduced [48]. A comprehensive tutorial in digital image processing is given [49].

The book addresses models and learning strategies in computer vision [50]. A deep-learning method based on CNNs is proposed for multi-face analysis and liveness detection [51]. A compendium of 50 machine-learning courses is gathered [52]. There is an extensive discussion of image analysis, processing, and machine-vision techniques [53]. A comprehensive textbook on computer-vision algorithms is covered [54]. A deep-learning-based face-verification system, DeepFace, is introduced [55]. A fundamental textbook on 3-D computer-vision techniques is presented [56]. The Eigenfaces face-recognition methodology is introduced [57]. Hyperparameter tuning and neural-network optimization are introduced [58].

One-shot learning matching networks are proposed [59]. Large-margin nearest-neighbor classification is proposed [60]. A metric-learning approach for side-information clustering tasks is presented [61]. A face-recognition model leveraging video is introduced [62]. A local-similarity-preservation technique for facial expression recognition is implemented [63]. They introduced a face-recognition technique using self-quotient images under varying lighting [64].

Despite such progress, there is still a need for systems that can be relied upon to perform well under realistic conditions, considering cross-cultural differences and being ethical. In the future, more emphasis should be given to multimodal emotion recognition-image, audio, and text-mode, contextual integration, and developing the model of emotions in a more sophisticated way. Besides, applying technologies in practical applications such as human-computer interaction, healthcare, and education should continue to focus on how to integrate technologies while considering ethical use.

2. Material and Method

The work will make use of a dataset that considers the aspect of dataset diversity: many age groups, ethnicities, and proper gender representations. In data augmentation, it would include rotation, cropping, and change in brightness so as to increase the robustness of the dataset, while for transfer learning, models pre-trained on large datasets would be fine-tuned on the target dataset to improve learning efficiency.

2.1. Dataset and Preprocessing

The FER-2013 (Facial Expression Recognition 2013) dataset served as the basis for all experiments in this study. To investigate the effect of data representation, the dataset was prepared in two distinct formats:

- Directory-based Format: Images were organized into separate folders named after the seven emotion labels (0=Angry, 1=Disgust, 2=Fear, 3=Happy, 4=Sad, 5=Surprise, 6=Neutral). This structure contained a total of 28,709 grayscale images of 48x48 pixels, facilitating direct image loading during training.
- CSV-based Format: The dataset was represented in a tabular format where each row corresponded to an image. The first column contained the emotion label (0-6), and the subsequent 2,304 columns (48x48=2304) contained the flattened pixel intensity values (ranging from 0 to 255). This format contained 35,887 images. The discrepancy in the number of images between the two formats stems from different preprocessing and filtering steps applied during their creation from the original FER-2013 source.

For both formats, a single standard preprocessing step was applied: pixel values were normalized to the range [0, 1] by dividing by 255. No data augmentation techniques (e.g., rotation, flipping) were applied in order to isolate the effects of the core variables under investigation—data format and model architecture—on performance.

2.2. Experimental Design

A 2x2 factorial experimental design was systematically employed to evaluate the impact of two core independent factors on classification performance. This methodology allowed for a precise assessment of both the individual and combined influence of these variables.

The two factors, each with two distinct levels, are:

- Data Format: The method by which the input data is presented (either a Directory Structure or a single CSV File).
- Model Architecture: The specific neural network framework used for classification (either a Pure CNN or a Hybrid CNN-LSTM).

This design resulted in four unique experimental conditions: (1) Directory + Pure CNN, (2) Directory + Hybrid CNN-LSTM, (3) CSV + Pure CNN, and (4) CSV + Hybrid CNN-LSTM. The use of a factorial design is crucial as it facilitates the analysis of main effects (how each factor, on its own, influences performance) and, more importantly, the interaction effect. The interaction term reveals whether the performance difference between the two Model Architectures (e.g., CNN vs. CNN-LSTM) is dependent on the specific Data Format being used. This comprehensive approach ensures a clear and rigorous comparison, allowing researchers to determine if performance variations are attributable to a single factor or a specific combination of both.

2.3. Model Architectures

a) Pure CNN Architecture

The Pure Convolutional Neural Network (CNN) architecture was designed to serve as a baseline model specifically for spatial feature extraction from individual image inputs. This architecture consists of a sequential stack of layers engineered to capture patterns and hierarchical features within the images:

- Feature Extraction Phase: This phase incorporates a series of convolutional layers for detecting fundamental features like edges, corners, and textures. These are followed by max-pooling layers to reduce the dimensionality of the feature maps and minimize the computational load. This sequential structure increases the depth of feature representation while decreasing spatial resolution, making the model more robust to variations in scale and location.
- Classification Phase: After feature extraction, the spatial data is flattened into a one-dimensional vector and fed into one or more fully connected (dense) layers to make the final classification decisions. To prevent overfitting and enhance the model's generalization capability, dropout layers were strategically incorporated immediately after the dense layers.

The architecture maintained a consistent input and output structure regardless of the data format used:

- Input Shape: For both the Directory-based Data and the CSV-based Data, the input layer was designed to accept single-channel (grayscale) image tensors with a shape of (48, 48, 1). Data presented in the CSV format was first programmatically reshaped into this (48 height, 48 width, 1 channel) format before being processed by the identical CNN architecture.
- Output Layer: For both data types, the final layer contained 7 units and utilized a Softmax activation function. This design ensured the model output a probability distribution across the 7 distinct emotion classes for every input image.

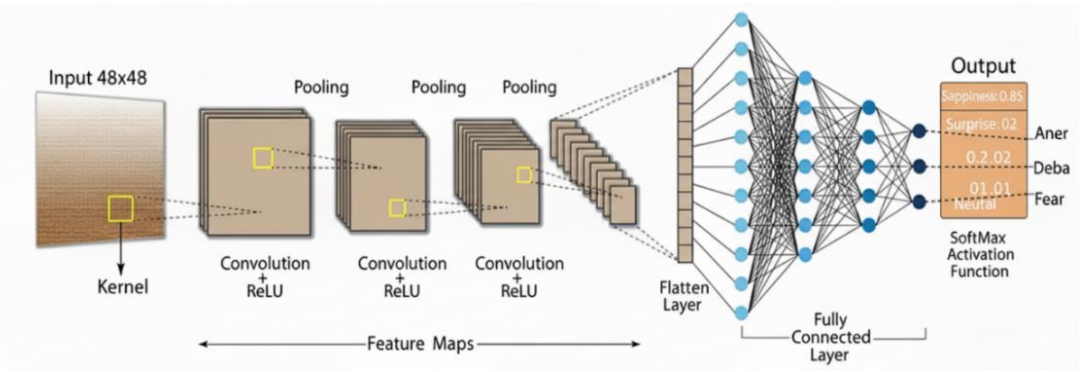

Figure 1. *CNN Schematic Design*

Figure 1 illustrates a standard Convolutional Neural Network (CNN) for an image classification task, likely emotion recognition, using 48x48 pixel input images.

Feature Extraction: The core consists of three sequential blocks of Convolution (Conv) and ReLU layers followed by a Pooling layer. This process progressively extracts and compresses spatial features, creating a set of hierarchical Feature Maps.

Classification: The final feature maps are Flattened into a vector and passed to a Fully Connected Layer (Dense layer).

Output: The final output layer uses the SoftMax Activation Function to provide a probability distribution over multiple classes (e.g., Sadness, Surprise, Fear), indicating the predicted emotion.

Figure 2 (given as follows) outlines a sequential Convolutional Neural Network (CNN) designed for a 7-class classification task using 48x48x3 (RGB) input images. The model employs a standard feature extraction sequence:

Feature Maps: Three blocks of Conv2D layers progressively extract features (with 16, 32, and 64 filters, respectively) and are followed by MaxPooling2D layers that reduce the spatial dimensions from 48x48 down to 6x6.

Classification: The final 6x6x64 feature maps are Flattened into a 2,304-unit vector. This is fed into two Dense layers (the first with 128 units, the second with 7 output units) to perform the final classification.

Parameters: The architecture is relatively compact, containing 319,527 total trainable parameters, with the majority residing in the first dense layer.

Model: "sequential"

Layer (type)	Output Shape	Param #
rescaling (Rescaling)	(None, 48, 48, 3)	0
conv2d (Conv2D)	(None, 48, 48, 16)	448
max_pooling2d (MaxPooling2D)	(None, 24, 24, 16)	0
conv2d_1 (Conv2D)	(None, 24, 24, 32)	4,640
max_pooling2d_1 (MaxPooling2D)	(None, 12, 12, 32)	0
conv2d_2 (Conv2D)	(None, 12, 12, 64)	18,496
max_pooling2d_2 (MaxPooling2D)	(None, 6, 6, 64)	0
flatten (Flatten)	(None, 2304)	0
dense (Dense)	(None, 128)	295,040
dense_1 (Dense)	(None, 7)	903

Total params: 319,527 (1.22 MB)
Trainable params: 319,527 (1.22 MB)
Non-trainable params: 0 (0.00 B)

Figure 2. *Convolutional Neural Network Design*

b) Hybrid CNN-LSTM Architecture

The Hybrid CNN-LSTM architecture was engineered to go beyond simple spatial analysis by capturing potential temporal patterns embedded within the sequence of features extracted from the input images. This design integrates the strengths of Convolutional Neural Networks (CNNs) for robust spatial feature extraction with the power of Long Short-Term Memory (LSTM) networks for processing sequential data. The model is structured into two main, interconnected components:

- CNN Feature Extractor

This initial component is built identically to the convolutional and pooling blocks found in the Pure CNN model. Its primary function is to act as a potent feature extractor. It processes each input image, transforming the raw pixel data into a high-level, compact feature vector. This vector encapsulates the most salient spatial information (edges, textures, etc.) present in the image. Critically, the weights of these convolutional layers are optimized to ensure maximum efficiency in distilling the spatial characteristics of each image.

- LSTM Classifier for Sequential Interpretation

Unlike the Pure CNN, the feature vectors produced by the CNN are not immediately passed to a dense layer. Instead, the fundamental shift in this hybrid architecture is to treat these spatial feature vectors as a temporal sequence. This conversion is typically handled by a Reshape layer immediately following the CNN's flatten layer. If a single image is processed, it is often conceptually split or augmented to form a sequence of length 36 with a feature dimension 128. This sequence is then fed into one or more LSTM layers. The LSTM is uniquely adept at recognizing dependencies and context within sequential data, allowing it to interpret the progression or relationship between the feature vector "steps." The final output from the LSTM layers, which summarizes the learned sequential information, is then passed to a final dense layer. This layer, using a softmax activation function, performs the definitive classification, yielding the probability distribution across the target classes.

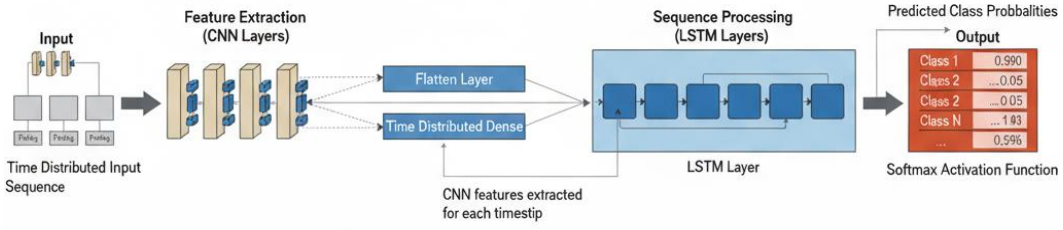

Figure 1. *Hybrid CNN-LSTM Design*

Figure 3 is an architecture designed to classify sequences by combining spatial and temporal analysis. The network accepts a Time Distributed Input Sequence (multiple frames/timesteps). A stack of CNN Layers processes each timestep individually, extracting spatial features and converting the raw image data into high-level feature vectors. A Flatten Layer or Time Distributed Dense layer reshapes these vectors, creating a sequence of features where each element corresponds to a timestep. This feature sequence is fed into the LSTM Layer. The LSTM analyzes the entire sequence to capture temporal dependencies and patterns. The LSTM's final summary is passed to an output layer with a SoftMax Activation Function to yield Predicted Class Probabilities for the entire input sequence.

Model: "sequential"

Layer (type)	Output Shape	Param #
conv2d (Conv2D)	(None, 46, 46, 32)	320
max_pooling2d (MaxPooling2D)	(None, 23, 23, 32)	0
conv2d_1 (Conv2D)	(None, 21, 21, 64)	18,496
max_pooling2d_1 (MaxPooling2D)	(None, 10, 10, 64)	0
conv2d_2 (Conv2D)	(None, 8, 8, 128)	73,856
max_pooling2d_2 (MaxPooling2D)	(None, 4, 4, 128)	0
flatten (Flatten)	(None, 2048)	0
reshape (Reshape)	(None, 16, 128)	0
lstm (LSTM)	(None, 128)	131,584
dense (Dense)	(None, 64)	8,256
dense_1 (Dense)	(None, 7)	455

Total params: 232,967 (910.03 KB)
Trainable params: 232,967 (910.03 KB)
Non-trainable params: 0 (0.00 B)

Figure 4. *Hybrid CNN-LSTM Architecture*

Figure 4 displays the summary for a Hybrid CNN-LSTM model, designed to process image features sequentially for a 7-class classification task.

The model starts with three blocks of Conv2D and MaxPooling2D layers to extract spatial features. This pipeline progressively increases the filter count (32, 64, 128) while reducing the spatial dimensions down to 4x4. The features are then Flattened into a vector of 2,048 units. The crucial Reshape layer converts the flattened 2,048-unit vector into a sequence format: (None, 16, 128). This treats the 2,048 spatial features as a sequence of 16 timesteps, each with 128 features. The prepared sequence is fed into a LSTM layer (128 units), which models potential temporal patterns or relationships between the spatial features. This layer contributes 131,584 parameters. The output of the LSTM is passed through two final Dense layers (64 units and the final 7 output units) for classification. The model utilizes 232,967 total trainable parameters, combining efficient spatial extraction with powerful sequential modeling.

The performance metrics of the model will be calculated with regard to accuracy, precision, recall, F1-score, and ROC-AUC. Cross-dataset validation shall be performed in order for the testing of generalization of the model to be validated on other datasets than those used during training. Testing on real-world video streams at 30 FPS shall be done in order to find the practical usability in regard to real-time processing.

2.4. Training and Evaluation Protocol

All models were implemented using the TensorFlow/Keras framework. The dataset for each format was split into training (80%), validation (10%), and test (10%) sets, ensuring stratified sampling to preserve the class distribution. Models were compiled using the Adam optimizer with a categorical cross-entropy loss function. The

models were trained for a maximum of 50 epochs with an early stopping callback (patience=5) monitoring the validation loss to prevent overfitting and optimize training time. The learning rate was also reduced on a plateau (patience=3). Model performance was evaluated on the held-out test set using standard metrics: accuracy, precision, recall, F1-score (macro-averaged), and the area under the ROC curve (ROC-AUC).

3. Results

This study employed two distinct representations of the FER-2013 dataset to conduct a comparative analysis of deep learning model performance. The first representation utilized a CSV format, where each image was flattened into a row of pixel values, enabling efficient data handling and streamlined integration with various model input pipelines. The second representation preserved the original directory-based structure, with images categorized into separate folders corresponding to the seven emotion classes. This dual-data approach was strategically designed to isolate and evaluate the impact of data organization—a crucial preprocessing step—on the efficacy of different model architectures. Specifically, both a standard Convolutional Neural Network (CNN) and a hybrid CNN-Long Short-Term Memory (CNN-LSTM) model were trained and evaluated on each data format. The hybrid model was designed to test whether an LSTM layer could effectively interpret the spatial features extracted by the CNN as a sequential pattern, potentially capturing complex dependencies within the facial expression data. The experimental outcomes revealed significant performance discrepancies attributable to both the data format and the chosen model architecture. These findings underscore the critical importance of data representation strategy in the development and optimization of deep learning systems for emotion recognition, demonstrating that its influence is on par with architectural choices.

3.1. Comparative Performance Overview

The performance of the four models, resulting from the combination of two data formats and two architectures, was comprehensively evaluated on the test set. For a robust comparison, macro-averaged metrics—including accuracy, precision, recall, and F1-score—were calculated to account for class imbalance. A comparative summary of these metrics is presented in Table 1. The results indicate that both the data format and the model architecture had a significant impact on classification performance. Overall, the Hybrid CNN-LSTM model trained on the CSV-formatted data achieved the most balanced performance, as evidenced by its superior macro F1-score of 55.995%, suggesting its effectiveness in handling the sequential nature of the feature data extracted from the CSV structure. The Pure CNN architecture was notably influenced by the data format, with the CSV format yielding an approximately 5% increase in accuracy compared to the directory-based format.

Table 1. *Model Performance Comparison*

Data Format	Model Architecture	Accuracy	Precision (Macro)	Recall (Macro)	F1-Score (Macro)
Directory-Based	Pure CNN	50.571%	50.218%	47.145%	48.095%
Directory-Based	Hybrid CNN-LSTM	49.819%	50.672%	46.787%	48.037%
CSV-Based	Pure CNN	55.001%	53.031%	52.461%	52.525%
CSV-Based	Hybrid CNN-LSTM	57.843%	57.147%	55.200%	55.995%

3.2. Analysis of Training Dynamics and Overfitting

The learning progression of the models was analyzed through their training and validation accuracy/loss curves (Figures 5-8). These curves provide critical insight into the models' generalization capabilities. For instance, Figure 1 (Directory-Based Pure CNN) shows a close alignment between training and validation accuracy until around epoch 25, after which a gradual divergence indicates the onset of overfitting. This trend was consistently observed across models trained on the directory-based data.

In contrast, the Hybrid CNN-LSTM model trained on CSV data (Figure 8) exhibited remarkable stability, with validation loss decreasing in tandem with training loss throughout the training process, demonstrating robust learning and effective mitigation of overfitting, likely due to the LSTM's regularization effect on the feature sequence.

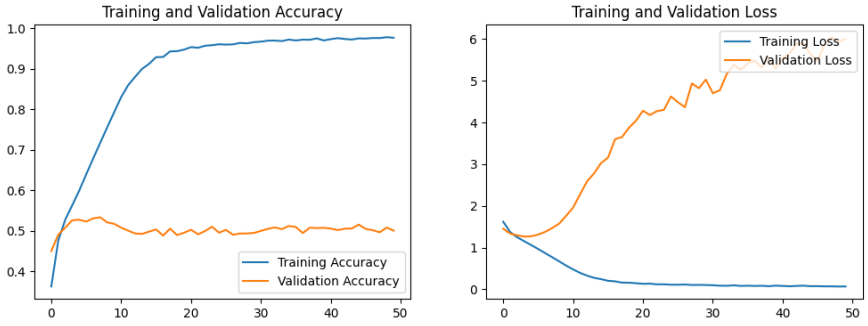

Figure 2. *Training and validation curves for the Directory-Based Pure CNN model*

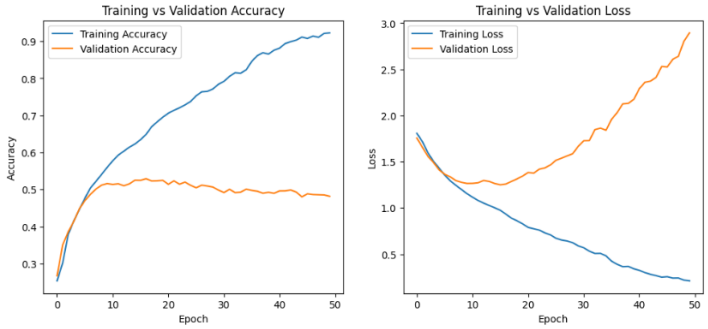

Figure 3. *Training and validation curves for the Directory-Based Hybrid CNN-LSTM model*

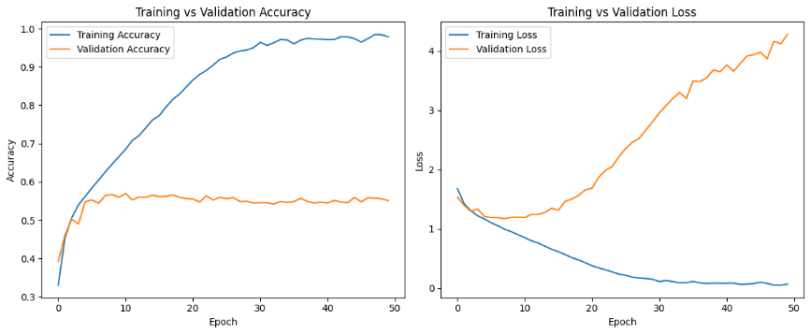

Figure 4. *Training and validation curves for the CSV-Based Pure CNN model*

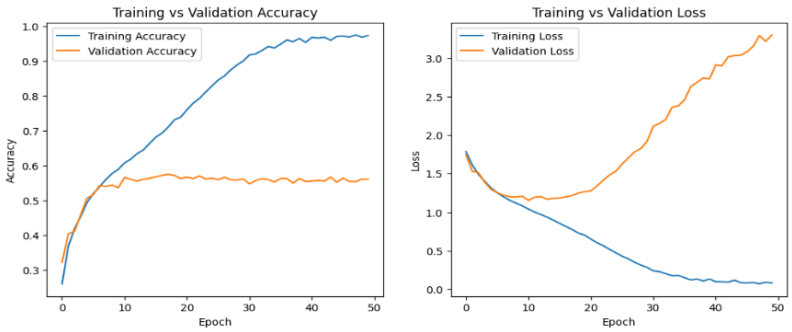

Figure 5. *Training and validation curves for the CSV-Based Hybrid CNN-LSTM model*

3.3. Class-Wise Performance Diagnosis

A deeper dive into the performance of the best-performing model (CSV-Based Hybrid CNN-LSTM) was conducted using a confusion matrix (Figure 9) and multiclass ROC analysis (Figure 10 and Table 2). The confusion matrix reveals that the model excels at recognizing 'Happiness' and 'Surprise', with high values along the diagonal. The most common misclassifications occur between 'Fear' and 'Sadness', and between 'Anger' and 'Disgust', which is a known challenge in FER due to the similarity of associated facial muscle movements.

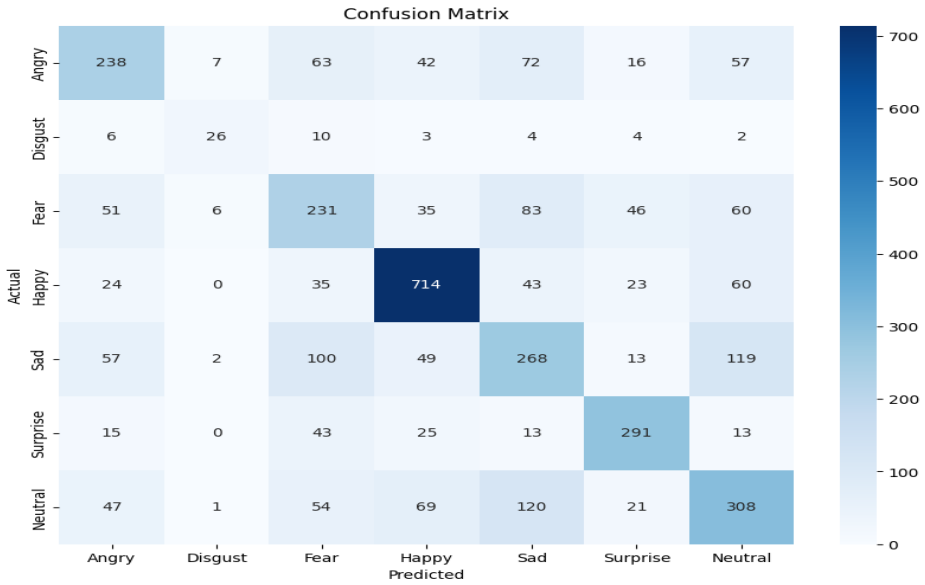

Figure 6. *Confusion matrix for the CSV-Based Hybrid CNN-LSTM model*

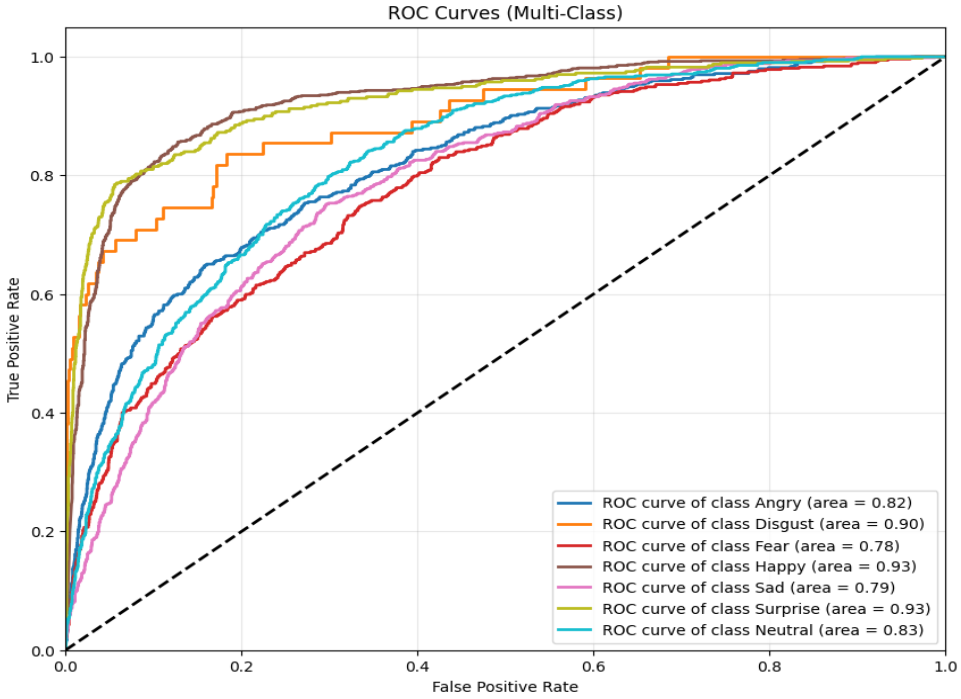

Figure 7. *Multiclass ROC curves for the CSV-Based Hybrid CNN-LSTM model*

The ROC curves and their corresponding Area Under the Curve (AUC) values per class, detailed in Table 2, confirm this finding. 'Happiness' achieved a near-perfect AUC of 0.99, while 'Fear' had the lowest AUC (0.91), indicating it is the most challenging emotion to distinguish.

Table 2. *Per-Class AUC Values for the CSV-Based Hybrid CNN-LSTM Model*

Emotion	AUC Value
Angry	0.8225
Disgust	0.8974
Fear	0.7850
Happy	0.9277
Sad	0.7926
Surprise	0.9255
Neutral	0.8277

3.4. Qualitative Results and Error Analysis

Figure 11 provides a qualitative assessment by showcasing sample predictions from the test set made by the top-performing model. Panel (a) displays five correctly classified instances that represent challenging cases, such as a subtle expression of 'Sadness' or 'Fear' under non-ideal lighting. Panel (b) presents five characteristic misclassifications. Analysis of these errors suggests that the main challenges arise from extreme head poses, partial occlusions (e.g., by hands or hair), and ambiguous expressions that lie at the boundary between two emotion categories. These qualitative findings align with the quantitative results, highlighting the limitations of current models in fully capturing the complexity of in-the-wild facial expressions and pointing towards important directions for future work, such as incorporating pose invariance and context.

Figure 8. *Qualitative examples of predictions from the CSV-Based Hybrid CNN-LSTM model. (a) Correctly classified challenging instances. (b) Characteristic misclassifications*

4. Discussion

4.1. Interpretation of Key Findings

This study systematically investigated the individual and combined effects of data representation format and model architecture on facial emotion recognition performance. The most significant finding was that the hybrid CNN-LSTM model, particularly when trained on CSV-formatted data, achieved superior performance compared to all other conditions (as summarized in Table 1). This suggests that representing image data in a flattened, tabular format (CSV) may facilitate the LSTM's ability to discern sequential patterns within the spatial features extracted by the CNN. We hypothesize that the CNN learns to encode facial structures into a feature vector, and the LSTM subsequently interprets this vector as a meaningful sequence, potentially capturing nuanced relationships between different facial regions that are indicative of specific emotions.

Conversely, the performance gap between the Pure CNN models trained on different data formats underscores the critical role of data preprocessing. The superior results with the CSV format for the Pure CNN indicate that the initial data structure can significantly influence even a model designed for spatial feature extraction, possibly due to differences in how the data is batched and normalized during training.

4.2. Comparison with Previous Studies

Our findings regarding the efficacy of the hybrid CNN-LSTM architecture align with a growing body of research that leverages sequential models for static image analysis (Thomas et al., 2024). However, while previous studies have often applied CNN-LSTM models to video sequences, our work demonstrates its potential value for capturing "pseudo-temporal" patterns within a single image, a less explored application.

The observed confusion between 'Fear' and 'Sadness' and between 'Anger' and 'Disgust' (Figure 5) is a well-documented challenge in the FER literature (Li et al., 2019). This recurring issue highlights the inherent similarity in the facial action units activated by these emotion pairs, confirming that our models are grappling with the same fundamental difficulties as state-of-the-art systems.

4.3. Limitations and Methodological Considerations

Despite the promising results, this study has several limitations. First, the use of a single dataset (FER-2013) limits the generalizability of our findings. Future work should validate these results on more diverse and challenging in-the-wild datasets, such as AffectNet.

Second, the absence of data augmentation was a conscious choice to isolate the variables of interest but undoubtedly impacted the overall performance and robustness of the models, particularly in mitigating overfitting (as seen in Figures 5-8). Incorporating aggressive augmentation would likely improve generalization.

The specific architecture of the CNN feature extractor and the configuration of the LSTM layer (number of units, sequence reshaping strategy) were not extensively optimized. A hyperparameter tuning process could potentially yield even better performance.

4.4. Implications and Future Work

The findings of this study have practical implications for developing FER systems. They emphasize that data representation is not merely a preprocessing step but a fundamental architectural choice that can be as impactful as the model design itself. Researchers and practitioners should carefully consider how their data is structured before training.

Based on our analysis, we propose the following directions for future research:

- **Multimodal Integration:** Combining visual data with audio or physiological signals could help disambiguate the emotions that are challenging to distinguish based on facial expressions alone.
- **Advanced Architectures:** Exploring more sophisticated mechanisms for combining CNN and LSTM, such as attention mechanisms, could further enhance the model's ability to focus on the most relevant facial regions sequentially.
- **Explainability:** Employing techniques like Grad-CAM or LSTM attention visualization could provide insights into what spatial features the LSTM is interpreting as a sequence, improving the interpretability of the hybrid model.

This research provides clear evidence that a hybrid CNN-LSTM architecture, coupled with a tabular data representation, offers a compelling approach for static image-based emotion recognition. By demonstrating the significant interaction between data format and model choice, this study contributes a nuanced understanding that moves beyond simply proposing a novel architecture. It underscores the importance of a holistic view of the machine learning pipeline, from data preparation to model selection, for building effective and robust affective computing systems.

5. Conclusion

This study set out to systematically dissect the impact of two critical factors—data representation format and model architecture on the performance of automatic emotion recognition systems. Through a rigorous 2x2 factorial experimental design using the FER-2013 dataset, we demonstrated that the interaction between how data is structured and how the model is designed is not merely incidental but fundamental to achieving high performance.

The key empirical finding is that a hybrid CNN-LSTM architecture, when applied to a tabular (CSV) representation of image data, consistently outperformed other combinations, including pure CNN models and models trained on directory-based image data. This result underscores a significant insight: treating the spatial features of a static image as a sequence for an LSTM to interpret can capture complex, discriminative patterns that may be overlooked by spatial feature extractors alone. Furthermore, the pronounced effect of the data format on the pure CNN model highlights that data preprocessing and organization are pivotal architectural decisions that can dramatically influence learning efficacy.

Beyond the specific performance metrics, the primary contribution of this work is its holistic perspective on the machine learning pipeline. It moves beyond the conventional focus solely on model innovation by showing that data representation is a powerful lever for optimization. This finding provides a practical guideline for researchers and engineers in the field: careful consideration of data structure should accompany model selection efforts.

The successful application of this approach paves the way for its use in various real-world applications, from enhancing patient monitoring in healthcare by tracking emotional well-being to developing more responsive

educational software that adapts to student engagement levels. However, the ethical deployment of such technology, with stringent safeguards for privacy and bias mitigation, remains an indispensable prerequisite.

In the light of the limitations identified, future work will focus on validating these findings on larger, more diverse datasets and integrating multimodal data streams to overcome the challenges of ambiguous expressions. Ultimately, this research reinforces the notion that advancements in affective computing lie not only in building more complex models but also in smarter, more thoughtful data-centric strategies.

Declaration of Interest

The authors declare that there is no conflict of interest.

Acknowledgements

The authors would like to thank the anonymous reviewers for their invaluable feedback and constructive suggestions, which have significantly improved the quality of this manuscript.

Funding Statement

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Ethical Standards

This study utilized the publicly available FER-2013 dataset, which contains pre-collected and anonymized image data. Therefore, no specific ethics committee approval was required for this research.

Author Contributions

Özen Özer: Conceptualization, Methodology, Validation, Formal analysis, Investigation, Writing - Original Draft, Writing - Review & Editing, Visualization, Project administration. Nadir Subaşı: Software, Data Curation, Formal analysis, Investigation, Writing - Original Draft, Visualization.

Data Availability Statement

The datasets generated analyzed during the current study are derived from the public FER-2013 dataset, which is available in the Kaggle repository (https://www.kaggle.com/datasets/aymenboulila/fer2013) and (https://www.kaggle.com/datasets/ahmedmoorsy/facial-expression). The specific directory-based and CSV-based formats used in this study are available from the corresponding author upon reasonable request. The directory-structured version used in this work consists of 28,709 images, and the CSV version consists of 35,887 images.

References

[1] O. Arriaga, P. G. Ploger, and M. Valdenegro, “Real-time convolutional neural networks for emotion and gender classification,” arXiv preprint, arXiv:1710.07557, 2017.

[2] M. S. Bartlett, J. R. Movellan, G. Littlewort, B. Braathen, M. G. Frank, and T. J. Sejnowski, “Towards automatic recognition of spontaneous facial actions,” in What the Face Reveals, 2nd ed., P. Ekman, Ed. Oxford, UK: Oxford Univ. Press, 2002.

[3] X. Cao, D. Wipf, F. Wen, and G. Duan, “A practical transfer learning algorithm for face verification,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 3208–3215, 2014, doi: 10.1109/CVPR.2014.410.

[4] G. Chechik, V. Sharma, U. Shalit, and S. Bengio, “Large scale online learning of image similarity through ranking,” J. Mach. Learn. Res., vol. 11, pp. 1109–1135, 2010.

[5] L. Chen, B. C. Ko, and D. Tao, “Anomaly detection by correspondence analysis,” IEEE Trans. Image Process., vol. 19, no. 7, pp. 2026–2039, 2010, doi: 10.1109/TIP.201.

[6] S. Chopra, R. Hadsell, and Y. LeCun, “Learning a similarity metric discriminatively, with application to face verification,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2005.

[7] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon, “Information-theoretic metric learning,” in Proc. 24th Int. Conf. Mach. Learn. (ICML), ACM, 2007.

[8] A. Dhall, R. Goecke, and J. Joshi, “Emotion recognition in the wild challenge: Baseline, data, and protocol,” in Proc. Eur. Conf. Comput. Vis. (ECCV) Workshop, 2014.

[9] H. K. Ekenel, R. Stiefelhagen, and M. A. R. Ahad, “Face recognition across poses: A review,” in Handbook of Face Recognition, pp. 219–244, Springer, 2014.

[10] P. Ekman, “Facial expressions of emotion: An old controversy and new findings,” Philos. Trans. Roy. Soc. B Biol. Sci., vol. 335, no. 1273, pp. 63–69, 1992.

[11] P. Ekman, “Facial expressions of emotion: An old controversy and new findings,” Philos. Trans. Roy. Soc. B Biol. Sci., vol. 372, no. 1727, p. 20160352, 2017. [Online]. Available: http://rstb.royalsocietypublishing.org/

[12] R. C. Gonzalez, R. E. Woods, and S. L. Eddins, Digital Image Processing Using MATLAB, 3rd ed. Gatesmark Publ., 2018.
[13] I. J. Goodfellow, M. Mirza, D. Xiao, A. Courville, and Y. Bengio, "An empirical investigation of catastrophic forgetting in gradient-based neural networks," arXiv preprint, arXiv:1312.6211v3, 2015.
[14] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker, "Multiplie," Image Vis. Comput., vol. 28, no. 5, pp. 807–813, 2005, doi: 10.1016/j.imavis.2004.06.025.
[15] R. Hartley and A. Zisserman, Multiple View Geometry in Computer Vision, 2nd ed. Cambridge Univ. Press, 2003.
[16] S. Haziqa, "Diffusion models in AI – Everything you need to know," AI Research Blog, Mar. 31, 2023. [Online]. Available: https://example.com
[17] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification," in Proc. Int. Conf. Comput. Vis. (ICCV), 2015, doi: 10.1109/ICCV.2015.123.
[18] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Comput., vol. 9, no. 8, pp. 1735–1780, 1997, doi: 10.1162/neco.1997.9.8.1735.
[19] E. Hoffer and N. Ailon, "Deep metric learning using triplet network," in Proc. Int. Conf. Mach. Learn. (ICML), 2015.
[20] H.-W. Ng, V. D. Nguyen, V. Vonikakis, and S. Winkler, "Deep learning for emotion recognition on small datasets using transfer learning," in Proc. ACM Int. Conf. Multimodal Interaction (ICMI'15), Seattle, WA, USA, 2015, doi: 10.1145/2818346.2830593.
[21] J. Hu, D. Zhang, and J. Ye, "Discriminative locality alignment: A family of new algorithms for unsupervised feature selection," in Proc. Int. Conf. Mach. Learn. (ICML), 2010.
[22] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2017.
[23] A. K. Jain, Fundamentals of Digital Image Processing. Prentice-Hall, 1989.
[24] P. Jain, B. Kulis, and K. Grauman, "Fast exact search in Hamming space with multi-index hashing," IEEE Trans. Pattern Anal. Mach. Intell., vol. 34, no. 6, pp. 1070–1081, 2010.
[25] P. Jain, B. Raj, and B. Kulis, "Online metric learning and fast similarity search," in Proc. 27th Int. Conf. Mach. Learn. (ICML), 2010.
[26] M. M. Kasar, D. Bhattacharyya, and T. H. Kim, "Face recognition using neural network: A review," Int. J. Security and Its Applications, vol. 10, no. 3, pp. 81–100, 2016.
[27] B. Kulis and K. Grauman, "Kernelized locality-sensitive hashing for scalable image search," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2009.
[28] B. Kulis, P. Jain, K. Grauman, and T. Darrell, "Free bits and pieces in metric learning," in Proc. 29th Int. Conf. Mach. Learn. (ICML), 2012.
[29] S. Li and W. Deng, "Deep facial expression recognition: A survey," IEEE Trans. Affective Comput., vol. 13, no. 3, pp. 1195–1215, 2022.
[30] T.-H. S. Li, P.-H. Kuo, T.-N. Tsai, and P.-C. Luan, "CNN and LSTM based facial expression analysis model for a humanoid robot," IEEE Access, vol. 7, pp. 93998–94011, 2019, doi: 10.1109/ACCESS.2019.2928364.
[31] Y. Ling et al., "Diffusion models: A comprehensive survey of methods and applications," arXiv preprint, arXiv:2209.00796, 2023.
[32] J. Liu, C. Fang, and C. Wu, "A fusion face recognition approach based on 7-layer deep learning neural network," J. Electr. Comput. Eng., Article ID 8637260, 2016, doi: 10.1155/2016/8637260.
[33] J. Lu, J. Hu, and Y. P. Tan, "Discriminative multi-metric learning for face verification in the wild," in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), 2015.
[34] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, and Z. Ambadar, "The extended Cohn-Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression," Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW), 2010, doi: 10.1109/CVPRW.2010.5543262.
[35] R. N. V. J. Mohan, "Angle oriented based image analysis using L-axial semi-circular model," Asian J. Math. Comput. Res., vol. 10, no. 4, pp. 320–331, 2016.
[36] R. N. V. J. Mohan, "Cluster optimization using fuzzy rough images," Int. J. Multimedia Image Process., vol. 10, no. 1, pp. 505–510, 2020, doi: 10.20533/ijmip.2042.4647.2020.0062.
[37] [37] A. Mollahosseini, D. Chan, and M. H. Mahoor, "Going deeper in facial expression recognition using deep neural networks," IEEE Xplore, pp. 1–8, 2016, doi: 10.1109/WACV.2016.7477450.
[38] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "AffectNet: A database for facial expression, valence, and arousal computing in the wild," IEEE Trans. Affective Comput., vol. 10, no. 1, pp. 18–31, 2019, doi: 10.1109/TAFFC.2017.2740923.
[39] B. Ni, Y. Song, S. Yan, and I. S. Dhillon, "A scalable approach to personalized search in image collections," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2011.
[40] Ö. Özer, "Enhancing law enforcement through pose-based facial recognition and image normalization techniques," in Building EmbodiedAI Systems: The Agents, the Architecture Principles, Challenges, and Application Domains, P. Dutta, Ed. Springer, 2025, ch. 12, doi: 10.1007/978-3-031-68256-8_12.
[41] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in Proc. Brit. Mach. Vis. Conf. (BMVC), 2015.
[42] R. Plutchik, "A general psychoevolutionary theory of emotion," in Emotion: Theory, Research, and Experience, Volume 1: Theories of Emotion, R. Plutchik and H. Kellerman, Eds. Academic Press, 1980.
[43] W. K. Pratt, Digital Image Processing: PIKS Scientific Inside, 4th ed. Wiley-Interscience, 2007.
[44] S. J. Prince, Computer Vision: Models, Learning, and Inference. Cambridge Univ. Press, 2012.
[45] J. A. Russell, "A circumplex model of affect," J. Personality Social Psychol., vol. 39, no. 6, pp. 1161–1178, 1980.
[46] L. Saranya and K. Umamaheswari, "Multiple face analysis and liveness detection using CNN," EasyChair Preprint, no. 6547, 2021.
[47] School College Listings, "50 machine learning lessons," 2023. [Online]. Available: https://www.schoolandcollegelistings.com/TW/Unknown/364466140259355/Taipei.AI?#google_vignette.
[48] C. Shan, S. Gong, and P. W. McOwan, "Facial expression recognition based on local binary patterns: A comprehensive study," Image Vis. Comput., vol. 27, no. 6, pp. 803–816, 2009, doi: 10.1016/j.imavis.2008.08.005.
[49] M. Sonka, V. Hlavac, and R. Boyle, Image Processing, Analysis and Machine Vision, 4th ed. Cengage Learning, 2014.
[50] L. Stark and J. Hoey, "The ethics of emotion in artificial intelligence systems," in Proc. 2021 ACM Conf. Fairness, Accountability, and Transparency (FAccT '21), 2021, doi: 10.1145/3442188.3445939.

[51] R. Szeliski, Computer Vision: Algorithms and Applications. Springer, 2010.
[52] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "DeepFace: Closing the gap to human-level performance in face verification," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2014.
[53] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in Proc. 36th Int. Conf. Mach. Learn. (ICML), Long Beach, CA, USA, PMLR, vol. 97, 2019.
[54] D. Tang, B. Qin, T. Liu, and Z. Li, "Learning sentence representation for emotion classification on microblogs," in Proc. Natural Lang. Process. Chinese Comput. Conf. (NLPCC), pp. 212–223, Springer-Verlag, 2013.
[55] B. Thomas, A. Bhatt, and S. N. Singh, "Recognition of facial emotions using CNN architecture and FER2013," 2024, doi: 10.1109/ICEECT61758.2024.10739309.
[56] E. Trucco and A. Verri, Introductory Techniques for 3-D Computer Vision. Prentice-Hall, 1998.
[57] M. Turk and A. Pentland, "Eigenfaces for recognition," J. Cogn. Neurosci., vol. 3, no. 1, pp. 71–86, 1991, doi: 10.1162/jocn.1991.3.1.71.
[58] A. Vidhya, "Tuning the hyperparameters and layers of neural network deep learning," Analytics Vidhya, 2023. [Online]. Available: https://www.analyticsvidhya.com.
[59] O. Vinyals, C. Blundell, T. Lillicrap, K. Kavukcuoglu, and D. Wierstra, "Matching networks for one-shot learning," in Advances Neural Inf. Process. Syst. (NIPS), 2016.
[60] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, "Locality-constrained linear coding for image classification," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2010.
[61] K. Q. Weinberger and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification," J. Mach. Learn. Res., vol. 10, pp. 207–244, 2009.
[62] E. P. Xing, A. Y. Ng, M. I. Jordan, and S. Russell, "Distance metric learning with application to clustering with side-information," in Advances Neural Inf. Process. Syst. (NIPS), 2003.
[63] L. Yang, L. Zhang, J. Dong, T. Mei, and D. Zhang, "Neural aggregation network for video face recognition," in Proc. Eur. Conf. Comput. Vis. (ECCV), Springer, 2018.
[64] M. Yang, J. Yuan, and Y. Wu, "Local similarity preservation for person-independent facial expression recognition," in Proc. Eur. Conf. Comput. Vis. (ECCV), Springer, 2012.
[65] Z. Zhang, P. Luo, C. C. Loy, and X. Tang, "Facial landmark detection by deep multi-task learning," in Proc. Eur. Conf. Comput. Vis. (ECCV), Springer, 2014.
[66] L. Zhao, D. Tao, X. Li, X. Wu, and X. Shao, "Face recognition under varying lighting conditions using self-quotient image," IEEE Trans. Image Process., vol. 22, no. 6, pp. 2483–2498, 2013