Indian Journal of Computer Science and Technology
https://www.doi.org/10.59256/indjcst.20260502057
Volume 5, Issue2 (May-August 2026), PP: 514-522.
www.indjcst.com
ISSN No: 2583-5300

Facial Emotion Recognition from Video Streams Using Deep Learning Techniques

Rubina S Pathan[1], Dr.Aslam J Karjagi[2]
[1] PG Scholar, Department of Computer Science and Engineering, Secab Institute of Engineering and Technology, Vijayapura, Karnataka, India.
[2]Associate.Professor, Department of Computer Science and Engineering, Secab Institute of Engineering and Technology Vijayapura, Karnataka, India.

To Cite this Article: *Rubina S Pathan[1], Dr.Aslam J Karjagi[2], "Facial Emotion Recognition from Video Streams Using Deep Learning Techniques", Indian Journal of Computer Science and Technology, Volume 05, Issue 02 (May-August 2026), PP: 514-522.*

Abstract: *Our Work address a deep learning-based approach on developing a deep learning-based system for automatic facial emotion recognition from video data. The proposed approach processes video input sequentially by extracting frames, detect-ing faces using the Multi-task Cascaded Convolutional Neural Network (MTCNN), and classifying each face into predefined emotion categories through a convolutional neural network model. A user-friendly web application built with Flask enables video upload, real-time inference, and extraction of segments corresponding to specific emotions. The model is trained and tested on the FER2013 dataset to ensure reliable performance. Experimental results indicate that the system effectively rec-ognizes facial expressions in video streams, achieving results comparable to established benchmarks. Additionally, a confusion matrix is utilized to evaluate classification performance across different emotion classes.*

Key Word: *Facial Emotion Recognition, Deep Learning, Convolutional Neural Network, MTCNN, Video Processing, Flask Web Application.*

I. INTRODUCTION

Facial emotion recognition has gained significant attention with the introduction of large-scale benchmark datasets such as FER2013, proposed by Goodfellow et al. [1], which enabled the development and evaluation of deep learning-based models for robust emotion classification. Facial expressions serve as a natural and effective means of conveying human emotions without the use of spoken language [2]. With the advancement of artificial intelligence, automatic facial emotion recognition (FER) has become an important research focus in recent years in the domains of computer vision and affective computing [3]. This technology has found applications in various areas including human–computer interaction, healthcare systems, education, surveillance, and entertainment. Deep learning ap-proaches have improved FER performance [1], [6], [9].

Recent progress in deep learning, particularly convolutional neural networks (CNNs), has greatly enhanced the capability of FER systems by enabling automatic extraction of complex features from visual data [1]. Despite these improvements, achieving reliable performance in real-world environments remains challenging due to variations in lighting conditions, head pose, occlusions, and differences in individual expres-sions [4].

Over time, FER approaches have evolved from traditional hand-crafted feature extraction methods to data-driven deep learning models that learn representations directly from images and video sequences. However, several limitations still persist, including sensitivity to low-quality inputs, class imbalance in datasets, and noisy annotations. These challenges highlight the need for more robust and adaptable emotion recognition frameworks.

To address these issues, modern systems incorporate ad-vanced face detection techniques such as Multi-task Cascaded Convolutional Networks (MTCNN) along with deep neural architectures for classification [5]. In this work, a video-based FER system is developed where input videos are processed frame by frame, faces are detected and aligned, and each face is classified into predefined emotion categories using a ResNet-18 model [6].

A. Computer Vision

Computer vision focuses on enabling machines to interpret and analyze visual information from images and videos. With the integration of deep learning, models are now capable of automatically learning meaningful features for tasks such as face detection, object recognition, and emotion analysis [1]. Facial emotion recognition represents a key application of computer vision, where the goal is to map facial patterns to corresponding emotional states.

B. Emotion Video Analysis

Unlike static image-based methods, video-based emotion analysis considers temporal variations in facial expressions, leading to improved recognition performance. This approach is particularly useful in applications such as driver monitoring,

adaptive user interfaces, behavioral analysis, and intelligent tu-toring systems. Continuous analysis of emotional cues enables systems to respond dynamically to user behavior.

C. Motivation and Objectives

The motivation behind this work arises from the increasing use of video data in real-world applications and the need for systems that can interpret human emotions over time. Temporal information present in videos provides valuable context that enhances recognition accuracy. The objectives of this study include designing an efficient deep learning-based framework for emotion detection, addressing challenges such as pose variation and noise, evaluating system performance using standard metrics, and implementing a user-friendly web interface for practical deployment.

II.RELATED WORK

Ian J. Goodfellow *et al.* introduced the FER2013 dataset, which has become a widely used benchmark for facial emo-tion recognition tasks [1]. The dataset consists of grayscale facial images collected from real-world scenarios and presents challenges such as low resolution, class imbalance, and noisy annotations.

Paul Ekman and Wallace V. Friesen introduced the Facial Action Coding System (FACS), which systematically cate-gorized facial expressions based on muscle movements [2]. This work laid the foundation for computational approaches to emotion analysis. Later, Rosalind W. Picard proposed the concept of affective computing, highlighting the importance of enabling machines to recognize and interpret human emotions [3].

Traditional machine learning methods relied on handcrafted features such as Local Binary Patterns (LBP) and Histogram of Oriented Gradients (HOG), but these approaches showed limited performance. The emergence of deep learning, par-ticularly convolutional neural networks (CNNs), significantly improved recognition accuracy.

Kaiming He *et al.* proposed the ResNet architecture, in-troducing residual learning to address the vanishing gradient problem in deep networks [6]. This architecture enables deeper networks and has been widely adopted for feature extraction in emotion recognition tasks.

Kaipeng Zhang *et al.* developed the Multi-task Cascaded Convolutional Network (MTCNN), which performs face de-tection and alignment simultaneously [5]. This method im-proves detection accuracy and ensures proper facial alignment before classification.

Christoph Pramerdorfer and Martin Kampel proposed a CNN-based approach combined with preprocessing and en-semble techniques to enhance recognition performance [9]. Although effective, this method increases computational com-plexity.

Omar Arriaga *et al.* introduced a lightweight CNN model designed for real-time emotion recognition [8]. The model achieves faster inference and is suitable for deployment in resource-constrained environments, though it may struggle with subtle expressions.

Ali Mollahosseini *et al.* presented large-scale datasets such as AffectNet, demonstrating that training on diverse data improves model robustness [10]. However, such approaches require significant computational resources.

Yu Li *et al.* proposed attention-based models that focus on important facial regions, improving recognition performance under occlusions [4]. Similarly, Trung Bui *et al.* addressed the issue of noisy annotations to enhance model reliability [15].

Ramprasaath R. Selvaraju *et al.* introduced Grad-CAM, a technique that improves model interpretability by visualizing important regions influencing predictions [20]. Additionally, Fulin Xue *et al.* and Kai Wang *et al.* proposed advanced methods to better capture complex emotional patterns [12], [14].

Despite significant progress, challenges such as dataset imbalance, low-resolution images, occlusion, and varying en-vironmental conditions remain. Furthermore, achieving a bal-ance between model accuracy and computational efficiency continues to be an important research problem.

Overall, deep learning-based approaches, particularly CNNs and residual networks, have significantly advanced facial emo-tion recognition. However, there is still a need for more robust and scalable systems for real-world applications.

III. OBJECTIVES

The primary objectives of this project include dataset prepa-ration, deep learning model design, performance evaluation using standard metrics, comparison with existing methods, and system deployment. FER2013 is used as the primary dataset, with extensive preprocessing and augmentation. A ResNet-18 backbone with transfer learning is adopted to balance accuracy and efficiency.

1) To design and implement an automated facial emotion recognition system capable of identifying human emo-tions from video data using deep learning techniques.
2) To develop an efficient video processing pipeline that performs frame extraction, face detection using MTCNN, and emotion classification using a convolu-tional neural network.
3) To train and evaluate the proposed deep learning model using the FER2013 dataset for reliable recognition of seven basic facial emotions.
4) To build a user-friendly web-based application using the Flask framework that enables video upload, emotion prediction, and emotion-specific video extraction.
5) To evaluate the performance of the system using stan-dard metrics such as accuracy, precision, recall, F1-score, and confusion matrix analysis.
6) To ensure real-time and scalable system performance while addressing challenges such as varying illumina-tion, multiple faces, and computational efficiency.

7) To analyze and visualize emotion-specific patterns in video data by generating emotion-filtered video sum-maries for improved interpretability and content under-standing.

IV.PROPOSED METHODOLOGY

A. System Architecture

The proposed text-query-based automatic emotion detection system employs a multi-component architecture that integrates deep learning, computer vision, and web technologies into a unified framework. The system follows a pipeline-oriented design with a clear separation of responsibilities across data acquisition, processing, and presentation layers, ensuring scal-ability and modularity.

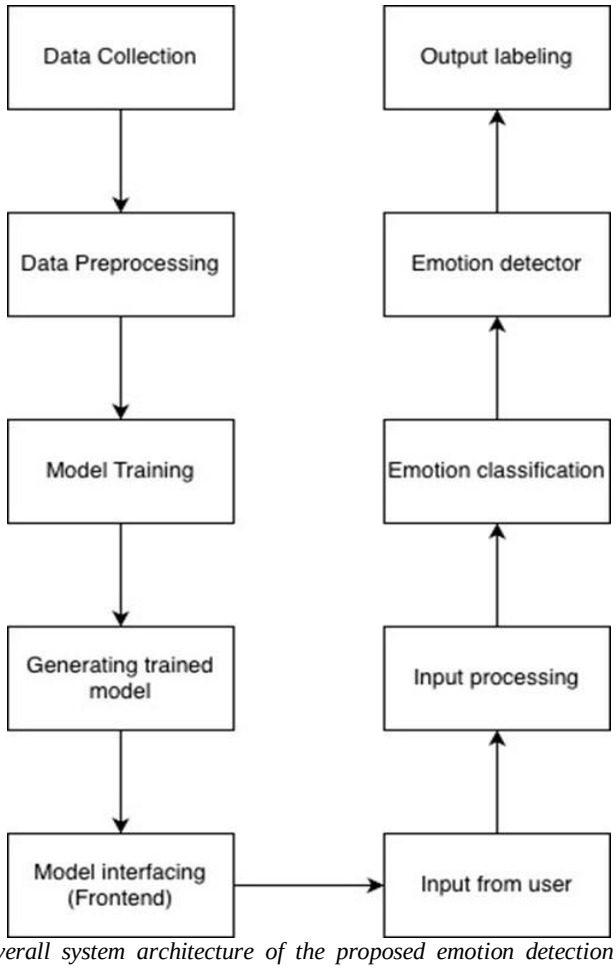

Fig. 1. Overall system architecture of the proposed emotion detection framework.

1) Architectural Overview: The overall architecture con-sists of the following major components:

1) **User Interface Layer (Frontend)**
- Web-based interface developed using HTML5, CSS3, and JavaScript
- Responsive design supporting both desktop and mo-bile platforms
- Real-time progress indicators and user feedback
- Interactive emotion visualization and video play-back

2) **Application Layer (Backend)**
- Flask web framework implemented in Python 3.11
- RESTful API endpoints for video upload and pro-cessing
- Session management for multiple concurrent users
- Asynchronous task execution

3) **Processing Layer (Machine Learning Pipeline)**
- Video frame extraction module
- Face detection using Multi-task Cascaded Convolu-tional Networks (MTCNN)
- Emotion classification using a ResNet18 model
- Temporal aggregation and prediction smoothing
- Annotated video generation

4) **Storage Layer**
- Uploaded video storage

- Cache for processed videos
- Emotion-specific video clips
- CSV files containing prediction logs
- Model checkpoints and configuration files

5) **Model Layer (Deep Learning)**

- Pre-trained ResNet18 backbone network
- Custom emotion classification head
- MTCNN face detection module
- GPU acceleration support

2) Detailed System Flow: The system workflow begins with video data acquisition from user uploads. Input data may contain noise and inconsistencies; therefore, preprocessing steps are applied to ensure uniform formatting and com-patibility with the trained model. During training, the deep learning architecture learns discriminative emotion patterns from labeled samples through iterative optimization.

After training, the learned weights and configurations are stored and used during inference. The trained model is integrated with the backend application, allowing users to interact with the emotion detection engine via the frontend interface. Uploaded videos are processed frame by frame, and predicted emotion labels are generated and displayed to the user.

3) Component Interaction Diagram: Figure 2 illustrates the interaction between system components, highlighting the flow of data across different layers and ensuring efficient communication between modules.

4) Data Flow Architecture: Input Data Flow:

User Video Upload → Session ID Generation (UUID) → Save to uploads/[session_id]_[filename].mp4→ Inference Trigger → Frame-by-frame Processing → Emotion Prediction and CSV Logging → Annotated Video Generation.

Processing Data Flow:

Video File → OpenCV Capture → Frame Extraction → RGB Conversion → MTCNN Face Detection → Face Cropping → Resize to 224 × 224 → Normalization → ResNet18 Inference → Softmax → Emotion Labeling.

Output Data Flow:

Processed Video → CSV Summary → Emotion Moment Extraction → Emotion-specific Video Generation → Playback and download

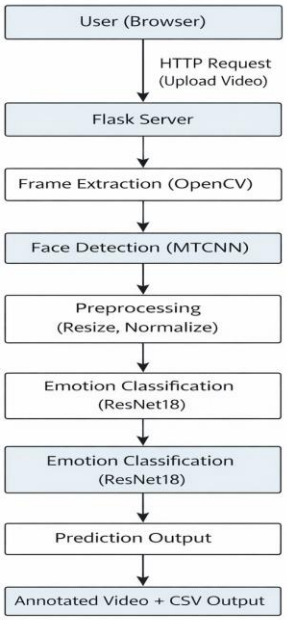

Fig. 2. Detailed data flow illustrating frame extraction, face detection, emotion classification, and output generation.

```
data/
├── Training/
│   ├── angry/ (3,995 images)
│   ├── disgust/ (436 images)
│   ├── fear/ (4,097 images)
│   ├── happy/ (7,215 images)
│   ├── sad/ (4,830 images)
│   ├── surprise/ (3,171 images)
│   └── neutral/ (4,965 images)
├── PublicTest/ (validation)
│   ├── angry/ (467 images)
│   ├── disgust/ (56 images)
│   └── ... (same structure)
└── PrivateTest/ (test)
    ├── angry/ (491 images)
    ├── disgust/ (55 images)
    └── ... (same structure)
```

Fig. 3. Sample output frame with detected face and predicted emotion label.

5) **Graphical Analysis:** To analyze the performance and training behavior of the model, graphical representations are used. Accuracy and loss curves over training epochs illus-trate model convergence, while a confusion matrix provides insight into classification performance across different emotion classes.

Emotion	Train	Validation	Test	Total	Percentage (%)
Angry	3995	467	491	4953	13.8
Disgust	436	56	55	547	1.5
Fear	4097	496	528	5121	14.3
Happy	7215	895	879	8989	25.0
Sad	4830	653	594	6077	16.9
Surprise	3171	415	416	4002	11.2
Neutral	4965	607	626	6198	17.3
Total	**28709**	**3589**	**3589**	**35887**	**100**

Table I EMOTION CLASS DISTRIBUTION IN THE FER2013 DATASET

B. Data Collection

1) **Dataset Source:** The Facial Expression Recognition 2013 (FER2013) dataset is used as the primary dataset for training and evaluation. It was introduced as part of the ICML 2013 Challenges in Representation Learning and is widely used as a benchmark in facial emotion recognition research.

Dataset Characteristics:

- Total images: 35,887 grayscale facial images
- Image resolution: 48 × 48 pixels
- Data format: CSV file containing pixel values and labels
- Annotation: Crowd-sourced labeling via Amazon Me-chanical Turk
- Publicly available dataset

2) **Emotion Class Distribution:**

3) **Dataset Limitations and Challenges:** Despite its widespread use, the FER2013 dataset presents several chal-lenges:

1) Label noise due to crowd-sourced annotations
2) Significant class imbalance, particularly for the disgust category
3) Low image resolution limiting fine-grained feature ex-traction
4) Absence of color information
5) Domain bias caused by web-scraped images
6) Lack of demographic and temporal metadata

4) **Mathematical Model:** Let x_i denote the detected face input and $f(\cdot)$ represent the deep neural network with param-eters ϑ. The feature representation is given by:

$$z = f(x_i; \vartheta) \quad (1)$$

The probability of each emotion class is computed using the Softmax function:

$$P(y = k \mid x_i) = \frac{e^{z_k}}{\sum_{i=1}^{C} e^{z_j}} \quad (2)$$

The predicted emotion is obtained as:

$$\hat{y} = \arg\max_k P(y = k \mid x_i) \quad (3)$$

Emotion	Train	Validation	Test	Total	Percentage (%)
Angry	3995	467	491	4953	13.8
Disgust	436	56	55	547	1.5
Fear	4097	496	528	5121	14.3
Happy	7215	895	879	8989	25.0
Sad	4830	653	594	6077	16.9
Surprise	3171	415	416	4002	11.2
Neutral	4965	607	626	6198	17.3
Total	**28709**	**3589**	**3589**	**35887**	**100**

TABLE II EMOTION CLASS DISTRIBUTION IN THE FER2013 DATASET

5) **Loss Function:** The model is trained using categorical cross-entropy loss, which measures the difference between predicted and actual emotion labels:

$$L = -\sum_{k=1}^{C} y_k \log(P(y = k \mid x_i)) \quad (4)$$

6) **Evaluation Metrics:** The performance of the proposed system is evaluated using standard classification metrics.

Accuracy:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (5)$$

Precision:

$$Precision = \frac{TP}{TP + FP} \quad (6)$$

Recall:

$$Recall = \frac{TP}{TP + FN} \quad (7)$$

F1-Score:

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \quad (8)$$

C. Data Collection

1) **Dataset Source:** The Facial Expression Recognition 2013 (FER2013) dataset is used as the primary dataset for training and evaluation. It was introduced as part of the ICML 2013 Challenges in Representation Learning and is widely used as a benchmark in facial emotion recognition research.

Dataset Characteristics:

- Total images: 35,887 grayscale facial images
- Image resolution: 48 × 48 pixels
- Data format: CSV file containing pixel values and labels
- Annotation: Crowd-sourced labeling via Amazon Me-chanical Turk
- Publicly available dataset

2) Emotion Class Distribution:
3) Dataset Limitations and Challenges*:* Despite its widespread use, the FER2013 dataset presents several chal-lenges:
1) Label noise due to crowd-sourced annotations
2) Significant class imbalance, particularly for the disgust category
3) Low image resolution limiting fine-grained feature ex-traction
4) Absence of color information
5) Domain bias caused by web-scraped images
6) Lack of demographic and temporal metadata

V.RESULTS AND DISCUSSION

This section presents qualitative and quantitative results of the proposed facial emotion detection system. The evaluation analyzes system performance, user interaction, and observed limitations. Sample outputs (Fig. 4) show detected faces with predicted emotion labels. The system performs reliably under normal lighting and accurately recognizes major emotions such as *Happy*, *Sad*, and *Surprise*. However, subtle expres-sions may lead to misclassification, with lower confidence for ambiguous emotions.

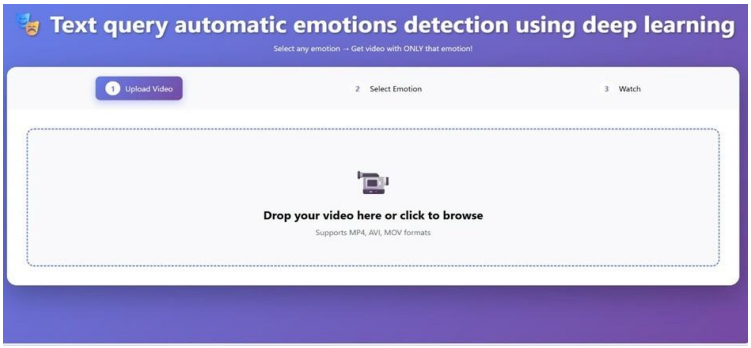

Fig. 4. Sample video frame with detected face and predicted emotion label

The web interface allows easy video upload (Fig. 5) and supports common formats such as MP4, AVI, and MOV.

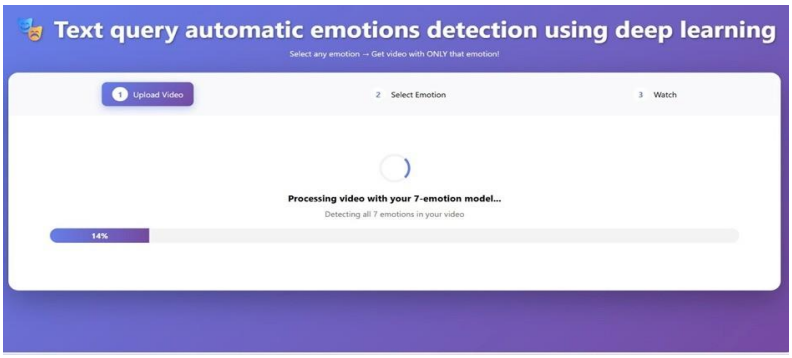

Fig. 5. Video upload interface of the application.

After uploading, the system processes the video with real-time progress display (Fig. ??).
In the sample video, *Happy* and *Neutral* emotions dominate, while others appear less frequently.

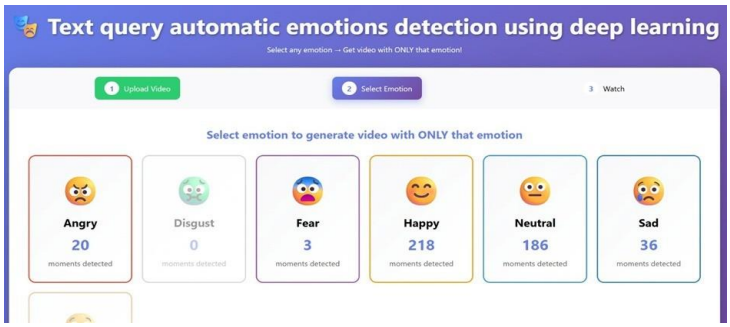

Fig. 6. Processing state showing real-time analysis

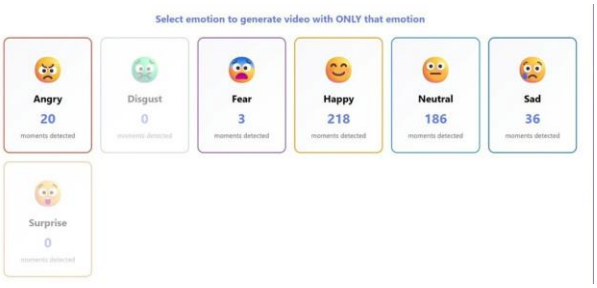

Fig. 7. Emotion summary of detected expressions

Fig. 8. Emotion selection interface.

The filtered output (Fig. 9) displays selected emotion frames.

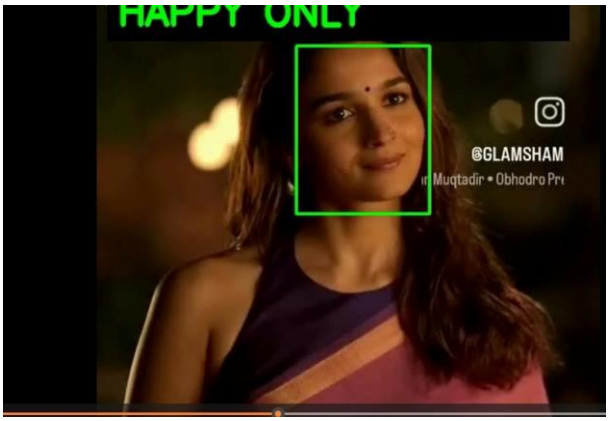

Fig. 9. Emotion-filtered video output.

1)Emotion Class Distribution: The distribution of emotion classes in the FER2013 dataset is illustrated in Fig. 10

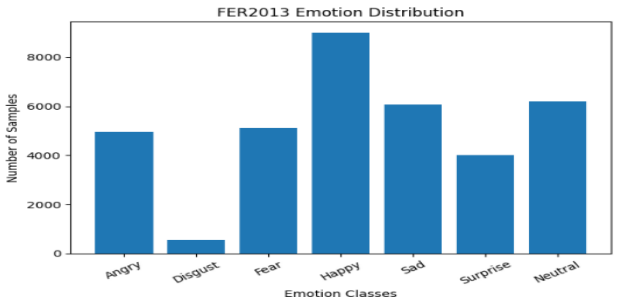

Fig. 10. Emotion class distribution in the FER2013 dataset

The dataset contains seven emotion categories, namely angry, disgust, fear, happy, sad, surprise, and neutral. It can be observed that the dataset is imbalanced, with the *happy* class having the highest number of samples, while the *disgust* class has significantly fewer samples.

Such class imbalance can affect the training process and may lead to biased predictions toward dominant classes. Therefore, appropriate preprocessing and training strategies are considered to mitigate this issue.

Final Remarks

The developed system shows that deep learning techniques can be effectively applied to real-time facial emotion analysis. The integration of detection, classification, and web-based interaction enables a practical solution for analyzing human expressions in video data. The system achieves reliable per-formance while maintaining usability for end users. However, certain challenges such as variations in lighting, pose, and subtle expressions still affect prediction accuracy. Future work will focus on improving model robustness, incorporating tem-poral features, and exploring multimodal inputs. Overall, the proposed system provides a solid basis for further development in emotion-aware computing applications.

VI.CONCLUSION

This work presents a video-based facial emotion recognition system built using deep learning techniques. The integration of face detection and classification enables efficient analysis of expressions across video frames. The system is supported by a web-based interface, allowing users to easily interact with the model and obtain results.

While the system performs well in identifying clear emo-tional states, challenges remain in handling complex scenarios involving subtle expressions and environmental variations. Future work will focus on improving model generalization, incorporating temporal information, and enhancing system performance for real-time applications.

Acknowledgement

The author sincerely acknowledges the guidance and sup-port provided throughout the development of this work. Ap-preciation is extended to the mentors and faculty members for their valuable suggestions and continuous encouragement.

The author is also thankful to the institution for offering the necessary infrastructure and resources required to complete this project successfully. Gratitude is further expressed to family and friends for their motivation and support during the entire process.

REFERENCES

1) I. J. Goodfellow *et al.*, "Challenges in representation learning: A report on three machine learning contests," *Neural Networks*, vol. 64, pp. 59–63, 2013.
2) P. Ekman and W. V. Friesen, *Facial Action Coding System*. Palo Alto, CA: Consulting Psychologists Press, 1978.
3) R. W. Picard, *Affective Computing*. MIT Press, 2000.
4) Y. Li *et al.*, "Occlusion-aware facial expression recognition using CNN with attention mechanism," *IEEE Trans. Image Processing*, vol. 28, no. 5, pp. 2439–2450, 2018.
5) K. Zhang *et al.*, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, 2016.
6) K. He *et al.*, "Deep residual learning for image recognition," in *Proc. IEEE CVPR*, pp. 770–778, 2016.
7) P. Lucey *et al.*, "The extended Cohn–Kanade dataset (CK+)," in *Proc. IEEE CVPR Workshops*, pp. 94–101, 2010.
8) O. Arriaga *et al.*, "Real-time convolutional neural networks for emotion and gender classification," *arXiv preprint arXiv:1710.07557*,2017.
9) C. Pramerdorfer and M. Kampel, "Facial expression recognition using convolutional neural networks," *arXiv preprint arXiv:1612.02903*, 2016.
10) A. Mollahosseini *et al.*, "AffectNet: A database for facial expression," *IEEE Trans. Affective Computing*, 2017.
11) L. Pham *et al.*, "Residual masking network," in *Proc. ICPR*, 2021.
12) F. Xue *et al.*, "Relation-aware facial expression recognition," in *Proc. ICCV*, 2021.
13) Z. Zhang *et al.*, "Interpersonal relation prediction," *IJCV*, 2020.
14) K. Wang *et al.*, "Suppressing uncertainties," in *Proc. CVPR*, 2021.
15) T. Bui *et al.*, "FER with noisy annotations," *IEEE Trans. Affective Computing*, 2022.
16) F. Schroff *et al.*, "FaceNet," in *Proc. CVPR*, 2015.
17) X. Li *et al.*, "Few-shot FER," in *Proc. AAAI*, 2022.
18) J. Deng *et al.*, "ImageNet," in *Proc. CVPR*, 2009.
19) L. Chen *et al.*, "Softmax regression for FER," *IEEE Trans. Affective Computing*, 2022.
20) R. Selvaraju *et al.*, "Grad-CAM," in *Proc. ICCV*, 2017.
21) S. Li *et al.*, "Crowdsourcing FER," in *Proc. CVPR*, 2017.
22) A. Mollahosseini *et al.*, "AffectNet extended," *IEEE Trans. Affective Computing*, 2019.
23) J. Russell, "Circumplex model of affect," *J. Personality*, 1980.
24) L. Barrett *et al.*, "Emotion perception," *Psychological Science*, 2011.
25) A. Paszke *et al.*, "PyTorch," *NeurIPS*, 2019.
26) G. Bradski, "OpenCV," *Dr. Dobb's Journal*, 2000.
27) M. Grinberg, *Flask Web Development*. 2018.
28) D. Merkel, "Docker," *Linux Journal*, 2014.
29) J. Grafsgaard *et al.*, "Emotion recognition in education," 2014.
30) McDuff *et al.*, "AM-FED dataset,"2013.