International Journal of Engineering Trends and Technology
ISSN: 2231–5381 / https://doi.org/10.14445/22315381/IJETT-V72I5P111

Volume 72 Issue 5, 105-112, May 2024

Original Article

Human Emotion Recognition System

Aharon Rushanyan[1], Artak Khemchyan[2]

[1,2]*Information Security and Software Engineering, National Polytechnic University of Armenia, Yerevan, Armenia.*

Corresponding Author : rushanyanaharon@gmail.com

Received: 18 January 2024 Revised: 10 April 2024 Accepted: 16 April 2024 Published: 26 May 2024

Abstract *- Emotions play an important role in human interaction and behaviour, affecting our decisions, interactions, and overall well-being. Facial expressions are a primary medium of conveying and understanding these emotions. According to David Mortensen's Communication Theory[1], only one-third of other people's emotions can be understood through words and tone of voice. At the same time, the remaining two-thirds come from facial expressions. expression (Mortensen, 2014). Understanding and recognizing emotions is an element in fields such as psychology, human-computer interaction, and artificial intelligence. Improving human-machine interaction involves recognizing and understanding human emotions. As a result, the field of emotion recognition technology has grown into a large industry, finding applications in various fields such as marketing research, driver impairment monitoring, user experience testing, and health evaluation [2]. In some cases, special schemes include human emotion recognition systems from video images used to identify facial expressions that allow the identification of basic human emotions.*

Keywords *- Emotion recognition, Video-based emotion recognition, Facial expressions, DeepFace library, Emotion detection technology.*

1. Introduction

Human emotion recognition systems are integral to fields as diverse as education, healthcare, business customer service, and entertainment. This study examines its performance through video-based emotion recognition, technical methodology, and associated challenges. Although significant progress has been made in emotion recognition technology, challenges remain, particularly in video-based recognition. Existing methodologies often struggle with different expressions influenced by lighting conditions, obstacles and individual differences. Additionally, the multidimensional nature of emotions requires a holistic approach, including facial expressions, vocalizations, body language, and contextual cues. In spite of much success, some real difficulties must be overcome, related to the totally automatically taken expressions when different lighting conditions occur, obstacles, postural changes and other individual differences. Strategies like facial expressions in the image or speech description in a video shoot will help in the student's comprehension.

This paper presents a new and promising video-based emotion recognition technique that leverages the strengths of what had been lacking in other methods. Different from typical previous studies which solely deal with facial expressions, this approach heavily uses a number of modalities, such as voices, icons and context sentences. By leveraging advanced technologies such as the DeepFace library [3] and employing sophisticated data processing and model development techniques, this research aims to enhance the accuracy and robustness of emotion recognition systems [16]. Furthermore, while existing research often overlooks the importance of data augmentation and extensive data labelling, this study emphasizes the importance of these steps in improving model generalization and performance in real-world scenarios. By meticulously curating diverse and well-annotated datasets such as CK+ [5], AffectNet [6, 17], and EmoReact [18], this model can efficiently recognize a wide range of emotional states in different settings.

However, emotional recognition cannot be compared to video, which constantly changes a person's emotions and creates a rich contextual background. In this study, video-based emotion recognition will be discussed, including its functionality, technical approach, and critical content. Detecting emotions in the video offers valuable insights into human behavior and emotional states. Only by studying facial expressions and their associated signals can it be possible to gain insight into a person's emotions, motives and reactions. Such knowledge is important in industries such as education, healthcare, commercial customer service and entertainment. Emotion recognition can create personalized learning experiences, improve mental health diagnoses [19], improve customer satisfaction, and create more engaging video games, among other applications.

However, the current literature lacks a comprehensive study of this challenge and the development of robust solutions. This study aims to solve this gap by looking at the technical complexity of performing emotion recognition through video, thus contributing to the development of human emotion recognition systems. Highlight the novelty and contribution of the work, comparing the approach with existing research findings. While previous studies have made progress in facial expression recognition, they have often neglected other mechanisms, such as the integration of tone and context cues. By incorporating these additional sources of information, this research provides a more comprehensive understanding of human emotions, leading to more accurate and robust emotion recognition systems.

Although human emotion recognition systems have made significant progress, they face some challenges and limitations. One of the major obstacles faced is managing emotions caused by factors such as lighting conditions, obstacles, changes in attitude and individual differences. These changes may affect the accuracy and reliability of the recognition system. Another limitation is that only analyzing expressions can capture the full range of emotional experiences. Emotions are complex and multidimensional, involving vocal sounds, body language, and contextual information rather than expressions. To mitigate this approaches such as reading facial expressions in images and interpreting speech in clips can be introduced to understand emotions [20] fully.

Overall, this research represents a significant advancement in the field of human emotion recognition, offering a holistic approach that considers multiple modalities and addresses key limitations of existing methodologies. Through rigorous experimentation and evaluation, to demonstrate the effectiveness and practicality of this approach, paving the way for enhanced human-computer interaction.

2. Methods

Several technologies or method-based approaches can be used to detect emotions in videos. There are many tools available, such as the DeepFace library [3], a Python-based library focused on emotion analysis and related applications. DeepFace provides an intuitive solution for face analysis with just a few lines of code for emotion recognition [4]. There are many advanced technologies and techniques that c can be used to show the exact emotion in your videos. This article will walk through the steps to create a human emotion detection system that is independent of external aids.

Following these steps will show how to create custom solutions for analyzing human emotions in different situations. Human emotion recognition systems from video images usually involve several steps to process and analyze facial expressions. First, facial features are used to identify key facial features such as eyes, nose, and mouth. The notes are used to analyze the landmarks later. Second, relevant emotional segments, including facial movements, shape changes, and texture patterns, are extracted from video frames. At the same time, interpret these characteristics by using machine learning algorithms to classify expressions into emotional aspects such as happiness, sadness, anger, fear, and surprise. Identifying emotions in videos is a step-by-step procedure. Let us go through each step in detail.

2.1. Data Collection

Data Collection: The foundation of a good emotion recognition system lies in the dataset used for training. These are aspects or phenomena of people in culture and events. It starts with a strongly labelled data set that accounts for this diversity.

Choosing the Right Database: Choosing the right dataset is very important to us. This choice was made considering a large number of emotions in different settings, for example, CK+ [5], AffectNet [6, 17] and EmoReact [18] databases.

Data Labeling: For accurate training, the database needs to be carefully annotated with appropriate emotional labels. Every emotion, in this case, happiness, sadness, anger and surprise, should be explained well.

Data Augmentation: Using data augmentation techniques to improve the robustness of the model. This method simulates different lighting conditions, angles, and facial expressions, creating changes to existing data. By increasing the database, this model can generalize to real-world scenarios.

Data Partitioning: It is very important to properly partition the data into three groups training, validation and test sets. This involves training the model on the training set by tuning the hyperparameters based on the validation set. Next, a test set will be used to evaluate how this model has performed so far. First, come down to collect good information to recognize emotions. Then, make sure the database is set up correctly for training.

2.2. Data Processing

Now that the curated database is prepared let us get ready to train the emotion recognition model with this data.

Data cleaning: Before running the study, remove outliers or noisy data points that may interfere with the performance of the model.

Image Processing (For Image-Based Recognition): If you are working with image data, preprocessing is important. All images were scaled consistently and converted to grayscale to simplify the model. Also, set the pixel value to fit the range {0, 1} or {-1, 1}.

Feature Extraction (For Voice-Based Recognition): In the case of voice-based emotion recognition, extract relevant voice features using techniques such as Melade Frequency Cepstral Coefficients (MFCCs) [7], pitch, and energy. These features serve as input to the model.

Label coding: Emotional labels were coded into numerical values. For example, 'happy' is 0, 'sad' is 1, etc. These metrics are important for training machine learning models.

Training-testing separation: Previously, the database was divided into training, cross-sectional and testing databases, usually between 80-10-10 or 70-15-15 split ratio.

Data Normalization: Data normalization includes scaling or normalization to ensure that each feature is at the same distance. That is, this step is important for some machine learning algorithms.

Data loading: To optimize the loading and processing of data in groups, various data pipelines were built. This is especially important when working with large data sets. Data processing plays an important role in compiling data that can be used to train emotion recognition models. When the initial data is ready, move on to the next step: model development.

2.3. Model Development

After filtering the data, it becomes possible to move on to developing and training an emotion recognition model.

Choosing the right model architecture: Choosing the right model architecture can have a significant impact on recognition accuracy. In this case, the choice falls on Convolutional Neural Networks (CNN) [8] for image-based recognition and Recurrent Neural Networks (RNN) [9] for time series data such as audio. CNNs are good at extracting spatial features from images, while RNNs are suitable for sequences such as audio data where temporal information is important.

Transfer Learning (optional): Using pre-trained models such as those provided by popular deep learning frameworks such as TensorFlow [10, 11] or PyTorch [29] can be a great time and resource saver. To take advantage of features learned from large databases, a model that is trained on a database was built. This approach helps speed up learning and can lead to better cognitive abilities.

Hyperparameter Tuning: Various hyperparameters must be tuned to achieve the best model performance—education level, party size, dropout rate, etc. Experiments were conducted to optimize parameters such as Fine-tuning this parameter is an important step in ensuring the accuracy of the model.

Selection of the growth function: In order to train the model effectively, it is necessary to choose an appropriate loss function. In multi-class classification problems were used categorical cross-entropy as it is often used in scenarios where data points belong to several classes. The loss function helps the model learn to predict the correct sentiment label for each input.

Model Training: Model training involves adjusting the weights and biases to minimize the chosen loss function and moving the grid forward and backward. This process replicates the model's ability to recognize emotions accurately. As training progresses, the model can learn how to map input data to emotional labels with more accurate predictions.

Equation for Softmax Activation: The Softmax activation function serves as an important element in emotion recognition. Convert raw model predictions to class probabilities so the results are clear. The Softmax equation is as follows:

$$\boldsymbol{P(yi \mid x) = \sum j = 1K\, ezi} \quad (1)$$

where,

the probability of class *i* given *x* is denoted by $P(yi \mid x)$.

zi is the uncorrected score. *i* score.

This means that there is a class *K* in general.

When it comes to the math behind how the probability is derived, knowing what the model means can help us better understand why a particular emotion is associated with a given set of input data and, after completing the model development phase, start another important step towards creating an autonomous human emotion detection system.

2.4. Model Evaluation

After training on the models, it is necessary to assess their reliability comprehensively.

Validation set performance: During training, the model's performance on the validation set was monitored for signs of overfitting or underfitting. The overshoot was prevented by stopping and checking the model in time.

Testing suite performance: After training, the model was tested on a test suite to estimate how it would perform in real life.

Metrics: The ability of the model to recognize different emotions was evaluated using appropriate evaluation metrics such as accuracy, precision, recall and F1 score.

2.5. Deployment

The model is now ready to deploy for practical use, as its performance was confirmed.

```
from flask import Flask, request, jsonify

import tensorflow as tf

app = Flask(__name__)

emotion_model = tf.keras.models.load_model("emotion_model.h5")

@app.route('/predict_emotion', methods=['POST'])

def prediction_emotion():
   data = request.json # Preprocess the data (e.g. convert the image to the correct format)

   predictions = emotion_model.predict(data) # Make predictions using the loaded model

 return jsonify({"predictions": forecastions.tolist()})# Post-processing and returning

results as JSON

if __name__ == '__main__':
   app.run(
```

Fig. 1 Create a Flask API endpoint

Web Service or API: To allow users and applications to access the model, need to be created a web service or API using Flask [13], FastAPI [14, 38], or Django [15, 37].

Code to create a Flask API endpoint: The following is an example of how you can create a Flask API endpoint for the emotion recognition model(Figure 1).

Security and access control: To ensure safe use of the emotion recognition service, authentication and access control mechanisms were implemented.

Monitoring and Logging: Monitoring and logging have been configured to monitor the performance and usage of the service, making it easier to detect problems and optimize the system. By following these steps, an automated human emotion recognition system was successfully developed, evaluated, and implemented.

2.5.1. Proposed System Architecture

A human emotion recognition system architecture is proposed to include a multi-modal framework that observes and analyses speech, images, and video frames with the aim of covering all details in real-time [21]. The architecture of this takes account of the latest fragment techniques, modular units and data processing systems with the aim of providing achievements in classifying complicated tasks in many applications [23].

3.1. Design Principles

3.1.1. Modularity

The architecture adopts a modular design approach, where each component is encapsulated and operates independently, facilitating flexibility, scalability, and maintainability [22]. Modular components include data collection modules, preprocessing modules, feature extraction modules, and model inference modules [24].

3.1.2. Extensibility

Within its design, the architecture can evanescently subsist as a warp space for new strategies deployment, fusion modalities, and algorithms. This allows user extensibility through well-defined interfaces and abstraction layers, which may be used to incorporate the features of new techniques and technologies without perturbing any other components of the system [25].

3.1.3. Adaptability

The framework is as flexible as possible and is modified in accordance with available data, environmental factors affecting the system, and application-related circumstances [26]. The system has the ability of adaptive mechanism, through parameter tuning and feedback loops, to alter its behavior and output according to current conditions in real-time [27].

3.1.4. Efficiency

Efficiency is a core requirement that runs through the essence of the architecture aimed at wringing the maximum out of resources and processing speed. Whether it be parallelization, asynchronous processing or distributed computing technology, adaptations are made to achieve this goal [23].

3.2. Components and Interactions

3.2.1. Data Collection Module

This section typically accumulates multi-modal data sources, including an audio stream, an image frame, and a video sequence. It is interactive with the outside sensors, the internal database, and the stream sources that provide it with real-time data inputs [28].

3.2.2. Preprocessing Module

The preprocessing module performs initial data cleaning, noise reduction, and normalization to enhance the quality of input data [16]. Techniques such as signal denoising, image enhancement, and feature scaling are applied to prepare the data for subsequent analysis [22].

3.2.3. Feature Extraction Module

Feature extraction techniques are employed to extract relevant features from the preprocessed data, capturing discriminative information for emotion recognition [24]. Feature extraction methods vary depending on the modality, including facial landmarks detection, speech feature extraction, and image feature extraction.

3.2.4. Fusion Module

The fusion module integrates features from multiple modalities and fusion strategies to enrich the representation of emotional cues [21].

Fusion techniques such as early fusion, late fusion, and multi-modal fusion are employed to combine information from different sources effectively [25].

3.2.5. Model Inference Module

The model inference module utilizes machine learning algorithms, deep neural networks, or statistical models to predict emotional states based on the fused feature representations [26]. Trained models are deployed within this module to perform real-time inference on input data streams [27].

3.3. Data Flow and Processing Pipeline

The data flow within the architecture follows a sequential processing pipeline, starting from data collection and ending with emotion prediction [19]. Extracted features from multiple modalities are fused using fusion strategies tailored to the specific task and application. The fused feature representation is fed into the emotion recognition model for inference, which outputs predicted emotional states. Feedback mechanisms and adaptive control loops may be incorporated to refine model predictions and adapt to changing conditions dynamically [29]—insights into Scalability, Flexibility, and Efficiency.

3.3.1. Scalability

The system architecture is set to be scaled out both vertically and horizontally, which allows expansion of the data volumes, computational loads, and user demands in line with their requirements [22]. Scalability is being gained by distributed processing, balancing distributed loads, and the strategies of resource allocation.

3.3.2. Flexibility

The architecture gives modality complementation and fusion functions flexible processing in terms of kinds of integration, the model's configuration and version [21]. The scientists can flexibly draw from the existing structure and borrow new techniques from elsewhere while doing so to meet the demand and still do their research in a new and extended way.

3.3.3. Efficiency

The architecture prioritizes efficiency in terms of computational resources, memory usage, and processing speed [26]. Techniques such as batch processing, caching, and model optimization are employed to maximize efficiency without compromising accuracy [27]. In summary, the proposed system architecture for human emotion recognition embodies advanced design principles, modular components, and efficient processing pipelines to enable accurate, scalable, and adaptive emotion analysis across multiple modalities. By integrating insights from diverse disciplines such as signal processing, machine learning, and human-computer interaction, the architecture offers a versatile framework for advancing the field of emotion recognition research.

4. CNNs and Confusion Matrix Analysis

In the context of emotion recognition, confusion matrices are valuable tools for assessing the performance of CNN-based models in classifying different emotional states [30]. CNNs are employed to process image data, particularly facial expressions, and predict the corresponding emotions. The output of the CNN model is a probability distribution over the emotion classes, indicating the likelihood of each emotion being present in the input image.

4.1. Prediction and Ground Truth Labels

For each input image in the evaluation dataset, the CNN model generates a predicted emotion label based on the highest probability output by the softmax layer. The respective ground truth labels, which represent the actual emotional states of the people in the pictures, are included in the dataset to test the model's outcomes with that which is actually observed [31].

4.2. Constructing the Confusion Matrix

The confusion matrix can then be developed based on tabulating the counts of true positive (TP), false positive (FP), true negative (TN), and false negative (FN) predictions for each emotion class. Within the framework of emotion detection, rows of the Confusion matrix represent ground truths, while columns-CNN model predictions [32].

4.3. Interpreting Confusion Matrix Entries

The entries of the confusion matrix indicate the model's performance in correctly classifying each emotion category. The diagonal elements (TP) represent the number of correctly predicted instances for each emotion class, indicating the model's ability to recognize those emotions accurately. Off-diagonal elements (FP and FN) represent misclassifications,

where the model incorrectly predicts one emotion as another. These entries provide insights into the model's weaknesses and areas for improvement [33].

4.4. Analyzing CNN Performance

By analyzing the distribution of entries in the confusion matrix, researchers can identify patterns of misclassification and assess the CNN model's strengths and limitations. Common patterns include confusion between similar emotions (e.g., sadness and fear) or underrepresentation of certain emotion classes due to imbalanced datasets. Metrics derived from the confusion matrix, such as precision, recall, and F1 score, provide quantitative measures of the CNN model's performance and help evaluate its effectiveness in recognizing different emotional states [34].

4.5. Iterative Model Improvement

Insights gained from confusion matrix analysis inform iterative model improvement strategies, such as fine-tuning CNN architectures, adjusting hyperparameters, or augmenting training data. By addressing the root causes of misclassifications identified in the confusion matrix, researchers can refine the CNN model's performance and enhance its accuracy in emotion recognition tasks.

5. Results

The model is now ready to deploy for practical use. Human emotion recognition systems have shown good performance, demonstrating the effectiveness of the adopted methodology. This section provides a comprehensive overview of the results, and further nuanced aspects are explored in the next section.

5.1. Model Performance Evaluation

A rigorous evaluation process was used to assess the effectiveness of the model. This section presents insights gained from evaluations during the training, validation, and testing phases.

5.1.1. Concept of Inspection

Continuous monitoring of model performance throughout the study reveals perceptions of redundancy or inadequacy. Adaptation strategies, including early termination, provide optimal results at this stage.

5.1.2. Performance Characteristics of the Test Set

After training, the model was rigorously tested in an independent database simulating real-world scenarios. This step determines the system's ability to generalize and identify emotions accurately in different contexts.

5.1.3. Comprehensive Criteria Assessment

A set of metrics, including accuracy, precision, recall, and F1 scores, were used to quantify model performance. Together, these metrics show the system's effectiveness in detecting different emotional states.

5.2. Smooth Deployment and Application Integration

From theoretical development to practical implementation marked the deployment phase. This section outlines the complex steps to ensure seamless availability and reliable use of emotion recognition models.

5.2.1. Export a Regular Model

Learning models are exported in versatile formats such as TensorFlow SavedModel or ONNX, facilitating effortless integration into various applications and frameworks.[12]

5.2.2. Intuitive Web Service/API Creation

To improve accessibility, intuitive web services or APIs are developed using frameworks such as Flask [13], FastAPI [14, 38], or Django [15, 37]. This allows users and applications to interact with the emotion recognition model seamlessly.

5.3. Anticipating Future Trends and Addressing Ethical Issues

This chapter considers emerging trends and ethical considerations in the field of human emotion recognition. It recognizes the evolving landscape of technology and its enormous impact on society. By presenting and discussing these unique findings, this paper makes a significant contribution to the ongoing conversation on human emotion recognition systems.

6. Advancements and Comparative Analysis

In comparison to state-of-the-art techniques reported in the literature, this approach to video-based emotion recognition yields superior results due to several key factors. In fact, this method covers more than just reading facial expressions, as it wants to detect all possible modes of non-verbal communication. The face-off expression is just one of the channels that are known; however, emotions are multidimensional and include external verbalization, body language, and other contextual factors [35]. This integrated way mobilizes the more sophisticated assessment of human emotions contributing in turn to the enhanced ability to detect them.

Secondly, these data collection and processing techniques are meticulously designed to ensure the robustness and generalization of the model. Diverse datasets that encompass a wide range of emotions were curated in various settings, leveraging datasets such as CK+ [5], AffectNet [6, 17], and EmoReact [18]. Through careful data labelling, augmentation, partitioning, and normalization, the training process was optimized, and the model's ability to generalize to real-world scenarios was improved.

Thirdly, this model development phase leverages cutting-edge deep learning architectures specifically tailored to handle different modalities of data. Convolutional Neural Networks

(CNN) for image-based recognition and Recurrent Neural Networks (RNN) for time series data such as audio have been used [24].

Additionally, transfer learning techniques have been used to capitalize on pre-trained models, accelerating the learning process and improving the model's cognitive abilities. Besides that, the internal methodology of this evaluation is so comprehensive that it involves not just training, validations and tests but everything in between as well. It was necessary to evaluate precision, recall precision, and F1 score to ensure the overall performance of the model [36].

This exhaustive evaluation validates the model's strong performance characteristics, including robustness and generalizability, such that the model has good coverage of diverse train and real-life scenarios. After that, this deployment method considers smooth integration and usability of the system to be the major features, thereby enabling the effective use of the emotion recognition system. This system implements interactive web services using APIs using platforms such as Flask [13], FastAPI [14, 38] or Django [15, 37], providing users with smooth and understandable scripts along with any application that can be used.

In summary, this approach to video-based emotion recognition outperforms existing techniques by adopting a holistic methodology, leveraging diverse and carefully curated datasets, employing state-of-the-art deep learning architectures, conducting rigorous model evaluation, and facilitating seamless deployment and integration. Through these strategic measures, superior results in recognizing and understanding human emotions have been achieved, paving the way for impactful applications across various domains.

7. Conclusion

The applications of human emotion recognition systems are vast and varied. In healthcare, these systems can be used to monitor and assess the emotional well-being of patients, helping in the diagnosis and treatment of mental health conditions. In education, emotion recognition technologies can improve teaching methods by adapting content and feedback based on students' emotional states and supporting personalized learning experiences. Additionally, in human-computer interaction, emotion-aware systems can enable more natural and intuitive interfaces, leading to improved user satisfaction and engagement. In addition to these domains, emotion recognition systems have promising applications in areas such as market research, security, and entertainment.

References

[1] C. David Mortensen, *Communication Theory*, Taylor & Francis, pp. 1-484, 2017. [Google Scholar] [Publisher Link]

[2] Don't Look Now: Why You Should Be Worried About Machines Reading Your Emotions, The Guardian, 2019. [Online]. Available: https://www.theguardian.com/technology/2019/mar/06/facial-recognition-software-emotional-science

[3] Yaniv Taigman et al., "DeepFace: Closing the Gap to Human-Level Performance in Face Verification," *2014 IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, OH, USA, pp. 1701-1708, 2014. [CrossRef] [Google Scholar] [Publisher Link]

[4] Boris Delovski, How Emotional Artificial Intelligence Can Improve Education, Edlitera, 2023. [Online]. Available. https://www.edlitera.com/blog/posts/emotional-artificial-intelligence-education

[5] Ali Mollahosseini, David Chan, and Mohammad H. Mahoor, "Going Deeper in Facial Expression Recognition using Deep Neural Networks," *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, Lake Placid, NY, USA, pp. 1-10, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[6] Mohammad H. Mahoor, AffectNet, Database. [Online]. Available: http://mohammadmahoor.com/affectnet/

[7] Sayf A. Majeed et al., "Mel Frequency Cepstral Coefficients (MFCC) Feature Extraction Enhancement in the Application of Speech Recognition: A Comparison Study," *Journal of Theoretical and Applied Information Technology*, vol. 79, no. 1, pp. 38-56, 2015. [Google Scholar] [Publisher Link]

[8] Karen Simonyan, and Andrew Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *arXiv*, pp. 1-14, 2014. [CrossRef] [Google Scholar] [Publisher Link]

[9] Kyunghyun Cho et al., "Learning Phrase Representations Using RNN Encoder-Decoder for Statistical Machine Translation," *arXiv*, pp. 1-15, 2014. [CrossRef] [Google Scholar] [Publisher Link]

[10] Martín Abadi et al., "TensorFlow: A Framework for Large-Scale Machine Learning," *12th USENIX Symposium on Operating System Design and Implementation (OSDI 16)*, pp. 265-283, 2016. [Google Scholar] [Publisher Link]

[11] Using the Saved Model Format, TensorFlow. [Online]. Available: https://www.tensorflow.org/guide/saved_model

[12] Open Neural Network Exchange (ONNX): Towards an Open Ecosystem for AI, Microsoft. [Online]. Available: https://github.com/onnx/onnx

[13] Flask Documentation, Palletsprojects. [Online]. Available: https://flask.palletsprojects.com/en/3.0.x/

[14] FastAPI is described as a Modern and High-Performance Web Framework for Developing APIs with Python 3.7+, Cilans System, [Online]. Available: https://cilans.net/ai-ml-python-r/fastapi-is-described-as-a-modern-and-high-performance-web-frameworkfor-developing-apis-with-python-3-7/

[15] Django Documentation, Django. [Online]. Available: https://docs.djangoproject.com/

[16] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton, "Deep Learning," *Nature*, vol. 521, no. 7553, pp. 436-444, 2015. [CrossRef] [Google Scholar] [Publisher Link]

[17] Ali Mollahosseini, Behzad Hasani, and Mohammad H. Mahoor, "AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild," *IEEE Transactions on Affective Computing*, vol. 10, no. 1, pp. 18-31, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[18] Steven R. Livingstone, and Frank A. Russo, "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A Dynamic, Multimodal Set of Facial and Vocal Expressions in North American English," *PLoS One*, vol. 13, no. 5, pp. 1-35, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[19] Jicheng Li et al., "MMASD: A Multimodal Dataset for Autism Intervention Analysis," *arxiv*, pp. 1-9, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[20] Tongshuai Song, Guanming Lu, and Jingjie Yan, "Emotion Recognition Based On Physiological Signals Using Convolutional Neural Networks," *Proceedings of the 2020 12th International Conference on Machine Learning and Computing*, pp. 161-165, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[21] Rosalind W. Picard, *Affective Computing*, MIT Press, pp. 1-292, 1997. [Google Scholar] [Publisher Link]

[22] Christopher M. Bishop, *Pattern Recognition and Machine Learning*, Springer New York, pp. 1-738, 2006. [Google Scholar] [Publisher Link]

[23] Xun Chen, Z. Jane Wang, and Martin McKeown, "Joint Blind Source Separation for Neurophysiological Data Analysis: Multiset and multimodal methods," *IEEE Signal Processing Magazine*, vol. 33, no. 3, pp. 86-107, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[24] Ian Goodfellow, Yoshua Bengio, and Aaron Courville, *Deep Learning*, MIT Press, pp. 1-800, 2016. [Google Scholar] [Publisher Link]

[25] Paul Ekman, "An Argument for Basic Emotions," *Cognition and Emotion*, vol. 6, no. 3-4, pp. 169-200, 1992. [CrossRef] [Google Scholar] [Publisher Link]

[26] Guoying Zhao, and Matti Pietikainen, "Dynamic Texture Recognition using Local Binary Patterns with an Application to Facial Expressions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 6, pp. 915-928, 2007. [CrossRef] [Google Scholar] [Publisher Link]

[27] Heysem Kaya, Furkan Gürpınar, Albert Ali Salah, "Video-Based Emotion Recognition in the Wild using Deep Transfer Learning and Score Fusion," *Image and Vision Computing*, vol. 65, pp. 66-75, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[28] Qingchen Zhang et al., "A Survey on Deep Learning for Big Data," *Information Fusion*, vol. 42, pp. 146-157, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[29] Adam Paszke et al., "PyTorch: An Imperative Style, High-Performance Deep Learning Library," *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*, pp. 8024-8035, 2019. [Google Scholar] [Publisher Link]

[30] Ziwei Liu et al., "Deep Learning Face Attributes in the Wild," *2015 IEEE International Conference on Computer Vision (ICCV)*, Santiago, Chile, pp. 3730-3738, 2015. [CrossRef] [Google Scholar] [Publisher Link]

[31] Emad Barsoum et al., "Training Deep Networks for Facial Expression Recognition with Crowd-Sourced Label Distribution," *Computer Vision and Pattern Recognition*, pp. 1-6, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[32] Zhongzheng Fu et al., "Emotion Recognition Based on Multi-Modal Physiological Signals and Transfer Learning," *Frontiers in Neuroscience*, vol. 16, pp. 1-15, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[33] Hailun Lian et al., "A Survey of Deep Learning-Based Multimodal Emotion Recognition: Speech, Text, and Face." *Entropy*, vol. 25, no. 10, pp. 1-33, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[34] Deepanway Ghosal et al., "DialogueGCN: A Graph Convolutional Neural Network for Emotion Recognition in Conversation." *arxiv*, pp. 1-11, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[35] Paul Ekman, *Emotions Revealed: Recognizing Faces and Feelings to Improve Communication and Emotional Life*, Henry Holt and Company, pp. 1-274, 2004. [Google Scholar] [Publisher Link]

[36] Marina Sokolova, and Guy Lapalme, "A Systematic Analysis of Performance Measures for Classification Tasks," *Information Processing & Management*, vol. 45, no. 4, pp. 427-437, 2009. [CrossRef] [Google Scholar] [Publisher Link]

[37] Daniel Roy Greenfeld, and Audrey Roy Greenfeld, *Two Scoops of Django 3.x: Best Practices for the Django Web Framework*, *Two Scoops Press*, 2019. [Publisher Link]

[38] T. Gutierrez, FastAPI Cookbook: Build Robust, Highly Scalable, and Secure Web APIs with Python, Packt Publishing, 2020. [Online]. Available: https://github.com/PacktPublishing/FastAPI-Cookbook