Inception-based Deep Learning Model for Arabic Audio Emotion Recognition in Forensics

Jaouhar Fattahi[1] ; Mohamed Mejri[1] ; Ridha Ghayoula[2] ; Sawssen Jalel[3];
Laila Boumlik[1] and Feriel Sghaier[4]

***Abstract*— Emotion recognition from audio signals is essential in forensic applications, offering insight into emotional states during interrogations, threat assessments, and crime scene analysis. This paper proposes an Inception-based deep learning model tailored for forensic arabic audio emotion recognition. The Inception architecture, with its multiscale feature extraction capabilities, efficiently captures subtle emotional details from complex audio signals. The model was evaluated on a dataset that represents a diverse range of emotional expressions, achieving superior performance in accuracy, robustness, and adaptability compared to traditional approaches. Its precision and ability to handle real-world variability make it particularly suited for forensic investigations. This work underscores the potential of advanced neural architectures in enhancing forensic decision-making and analysis.**

***Index Terms*— Audio, Emotion, Recognition, Deep Learning, Inception, Forensics, Cybersecurity.**

I. Introduction

Forensic science essentially revolves around gathering evidence in that it could be analyzed to reveal the truth behind various incidents such as physical crimes or digital intrusions, in addition to other disputes that arise. Our era has seen a growing need of digital forensics that specifically deals with investigating criminal activities and security breaches occurring in the cyberspace. Forensics encompass tasks such as examining drives for data retrieval and determining cyberattack origins as well as identifying those responsible, for such actions. Its significance has increased in tandem with the increasing dependence on technologies. The relationship between cybersecurity and forensics [1]–[7] is closely connected as they work hand in hand to safeguard systems and data from intrusions or breaches in security measures. When cyberdefenses fail, digital forensics steps in to uncover details about cyberincidents aiding organizations and law enforcement agencies in unraveling the what, how, and who are behind events. Recognizing emotions is one of prime important focuses in forensic investigations. It provides experts and authorities with better insights on a person's intentions, honesty, or state of mind. For instance, amid interrogations, picking up on emotions like stress, anxiety, or calmness can help investigators figure out whether someone is lying or telling the truth. In the same vein, tracking emotions [8], [9] in threat assessments—like detecting anger, fear, or distress in a phone call or a surveillance recording—can provide valuable clues for solving cases. Emotion recognition also comes in handy when working with victims of crimes. By analyzing how someone speaks, experts can pick up on signs of trauma and provide better support tailored to what the person is going through. In recent years, deep learning has completely changed how we approach emotion recognition. Instead of relying on manually crafted methods, these models can dig into audio recordings and automatically pull out patterns and details that people might miss. This has massively boosted accuracy, especially in tough scenarios like forensic cases, where audio quality isn't always perfect. With neural networks like convolutional and recurrent models, it is now possible to sort through large amounts of audio and figure out emotional clues from different kinds of expressions or environments. One model that really stands out for this job is the Inception architecture. It first showed up in 2014 as part of Google's GoogLeNet [10], [11] for image classification and has been praised for its outstanding way of working. Inception processes data at different scales at the same time, which makes it adequate at picking up on both small details and bigger patterns. Even though it was originally built for image-related tasks, researchers have found ways to make it work for audio by using spectrograms—visual representations of sound—as images, or just numerical inputs. In fact, it has already been put to good use in things like speech recognition and music analysis, proving how adaptable and reliable it is. With its ability to handle complicated data efficiently, Inception is in many cases a perfect fit for forensic audio emotion recognition, where getting things right really matters. Finally, being aware that every language has its particularities, we precise that our study covers the Arabic audios only. In this paper, we propose an Inception-based model that focuses on recognizing Arabic emotions from audio in forensic settings. We will walk through how the model works, test it on a dataset with diverse emotional expressions, and compare it with traditional methods to show why it could be an adequate solution for real-world forensic challenges.

[1] Jaouhar Fattahi and Mohamed Mejri and Laila Boumlik are with the Department of Computer Science and Software Engineering, Laval University, Quebec, Canada. `Jaouhar.Fattahi.1@ulaval.ca; Mohamed.Mejri@ift.ulaval.ca; Laila.Boumlik.1@ulaval.ca`

[2] Ridha Ghayoula is with the Faculty of Engineering, University of Moncton, New Brunswick, Canada. `Ridha.Ghayoula@umoncton.ca`

[3] Sawssen Jalel is with TEKUP University, Tunis, Tunisia. `Sawssen.Jalel@tek-up.tn`

[4] Feriel Sghaier is with the Carthage National Engineering School, Carthage University, Tunis, Tunisia. `Feriel.Sghaier@enicar.ucar.tn`

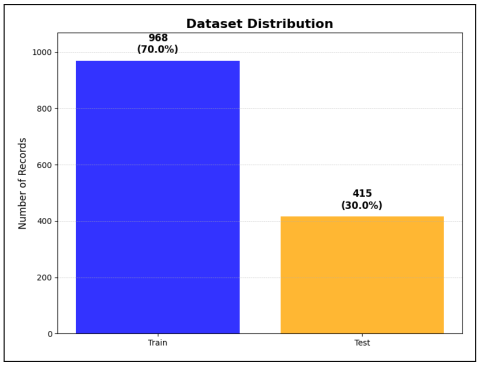


Fig. 1: Dataset distribution

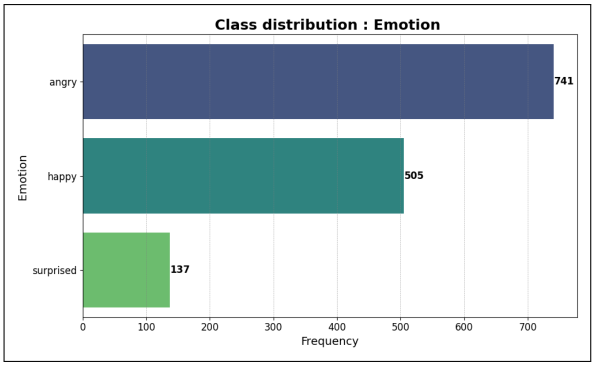


Fig. 2: Emotion classes

II. Dataset

In our study, we have used the Arabic Natural Audio Dataset (ANAD) [12]. The dataset comprises 1384 speech records extracted from eight Arabic talk show videos featuring live calls between an anchor and a caller. Each video was segmented into turns for the anchor and the caller. Eighteen listeners labeled the emotional content of each turn as either happy, angry, or surprised. To ensure quality, silences, laughter, and noisy segments were excluded, and the remaining speech data was divided into 1-second units. A total of 25 low-level acoustic features, including intensity, zero crossing rate, MFCC 1-12, F0 (fundamental frequency), F0 envelope, probability of voicing, and LSP frequency 0-7, were extracted from the audio. For each feature, 19 statistical functions such as mean, maximum, range, standard deviation, and skewness were computed, along with delta coefficients as first derivatives, resulting in a feature set of 950 attributes, among them only 844 attributes are used in our experiment. Table I shows the audio features and methods to extract them from WAV files of the dataset. Table II shows the statistical functions applied to audio features. Table III shows the dynamic features. Table IV shows the overall features. Fig. 1 shows the data distribution between training and testing. Fig. 2 presents the distribution of emotion classes over the dataset. Fig. 3 visualizes a sample of features extracted from an audio wave file in the dataset using the library Librosa.

TABLE I: Audio features and methods to extract them from a WAV File

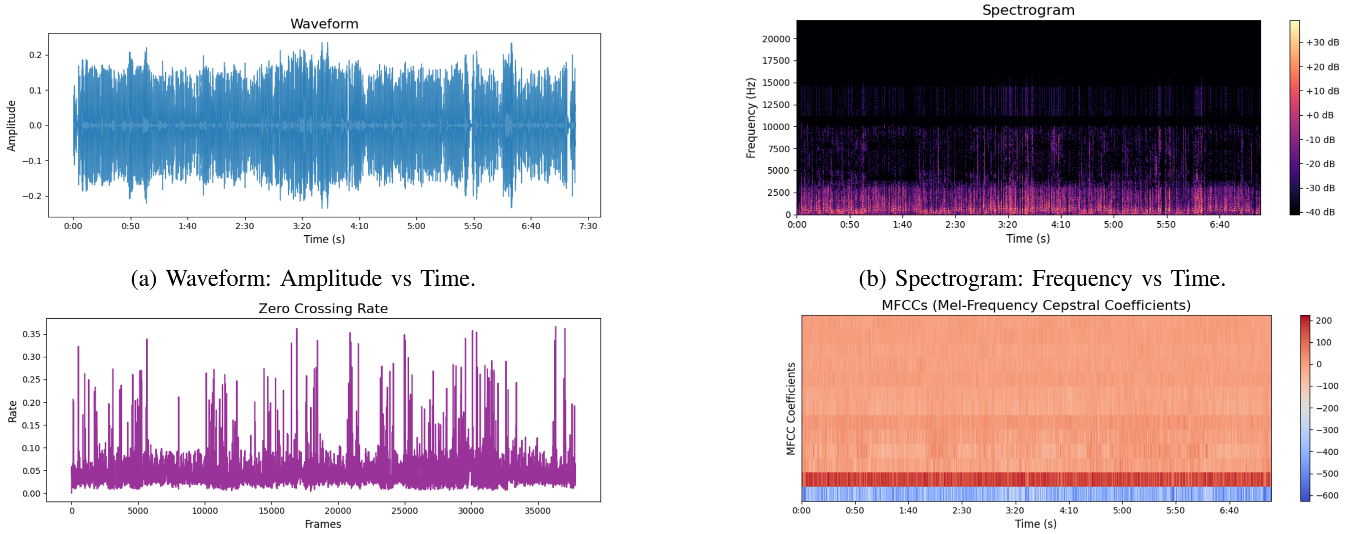
Feature	Definition	Extraction Method
Intensity	Represents the perceived loudness of the audio signal, often related to its amplitude.	Use libraries like LibROSA [13] to calculate amplitude envelopes.
Zero crossing rate	The rate at which the audio waveform crosses the zero amplitude line, useful for identifying tonal vs. percussive sounds.	Extract using LibROSA's `zero_crossings()` function.
MFCC 1-12	Mel-frequency cepstral coefficients; represent the short-term power spectrum of sound, capturing timbral features.	Compute with `librosa.feature.mfcc()` in LibROSA.
F0 (Fundamental frequency)	The lowest frequency of a periodic waveform, corresponding to the pitch of the sound.	Extract using tools like parselmouth [14] or Praat [15].
F0 envelope	A smoothed curve outlining the variations in the fundamental frequency over time.	Use pitch tracking algorithms like `pYIN` in LibROSA.
Probability of voicing	The likelihood that a segment of audio corresponds to a voiced sound (e.g., vowels).	Compute with pitch estimation libraries such as Parselmouth [14].
LSP frequency 0-7	Line Spectral Pairs (LSP); represent spectral properties of speech, often used for compression and synthesis.	Use specialized tools like SpeechPy [16] or Kaldi [17].

TABLE II: Statistical functions applied to audio features

Statistical Function	Definition
Maximum	The highest value of the feature over time.
Minimum	The lowest value of the feature over time.
Range	The difference between the maximum and minimum values.
Absolute Position of Maximum	The time index of the maximum value.
Absolute Position of Minimum	The time index of the minimum value.
Arithmetic Mean	The average value of the feature over time.
Linear Regression 1, 2, A, Q	Regression coefficients representing trends or slopes in the feature values.
Standard Deviation	A measure of variability or dispersion in the feature values.
Kurtosis	The ”peakedness” of the feature value distribution.
Skewness	The asymmetry of the feature value distribution.
Quartiles (1, 2, 3)	Values dividing the data into quarters.
Inter-Quartile Ranges (1-2, 2-3, 1-3)	Differences between quartiles, representing variability.

TABLE III: Dynamic features

Dynamic Feature	Definition
Delta Coefficient	An estimate of the first derivative of each low-level descriptor (LLD), capturing dynamic changes over time.



(a) Waveform: Amplitude vs Time.

(b) Spectrogram: Frequency vs Time.

(c) Zero Crossing Rate: Rate vs Frames.

(d) MFCCs: Coefficients vs Time.

Fig. 3: Visualization of a sample of features extracted from an audio wave file in the dataset using Librosa.

TABLE IV: Summary of audio features

Descriptor Type	Count	Details
Low-Level Descriptors (LLDs)	25	Intensity, zero crossing rate, MFCC 1-12, F0, F0 envelope, probability of voicing, LSP frequency 0-7.
Statistical Functions	19	Applied to each LLD to compute key properties.
Dynamic Features (Delta)	25	Captures temporal changes for each LLD.
Total Features	950	$25 \times 19 + 25 = 950$.

TABLE V: Summary of audio features

III. Model description

Our deep learning is based on the Inception v1 architecture. It is designed for recognition of three classes: *angry*, *happy*, and *surprised*. The Inception module extracts multi-scale features by processing the input data through four parallel branches:

- a 1×1 convolution,
- a 1×1 convolution followed by a 3×3 convolution,
- a 1×1 convolution followed by a 5×5 convolution, and
- a max-pooling layer followed by a 1×1 convolution.

These branches capture fine-grained and broader patterns in the input audio features, which are reshaped into a 3D tensor to include the channel dimension.

The model consists of two stacked Inception modules, each followed by a max-pooling layer to reduce dimensionality and retain important features. The output of the final module is flattened and passed through dense layers, including:

- a dropout layer for regularization, and
- a fully connected layer with 128 neurons for further feature extraction.

The final output layer uses a *softmax* activation function to classify the audio into one of the three emotion classes.

The model is trained using the *Adam optimizer* and sparse categorical cross-entropy loss over 40 epochs with a batch size of 16. Fig. 4 presents the overall model.

IV. Results

Fig. 5 and Fig. 6 shows, respectively, the evolution of the overall accuracy and loss of our model over epochs (40 epochs used). Fig. 8 shows the confusion matrix of our model. Fig. 7 shows the performance of our model. As we can see it in Fig. 7, regarding the first class (angry) our model reached 97% of accuracy, 96% of precision, 97% of recall and 96% of F1-score. Regarding the second class (happy), our model reached 91% of accuracy, 99% of precision, 91% of recall and 95% of F1-score. With respect to the third class (surprised), our model reached 89% of accuracy, 77% of precision, 82% of recall and 95% of F1-score. The overall performance of our model is very satisfactory, which positions it as a good candidate for emotion recognition.

V. Discussion and comparison with related work

Emotion recognition, particularly leveraging audio and multimodal approaches, has seen significant advancements with the integration of deep learning and novel frameworks. Here, we review several recent contributions in the field, highlighting their methodologies and applications. Jin and Zai [18] proposed an audiovisual emotion recognition model based on a bi-layer LSTM and multi-head attention mechanism. Using the RAVDESS dataset, the authors demonstrated that combining temporal modeling with attention mechanisms significantly improves emotion recognition accuracy by capturing dependencies between audiovisual signals. Their work highlights the effectiveness of hybrid architectures in processing multimodal inputs. Raheel [19] focused on emotion recognition in a 3D space, using EEG

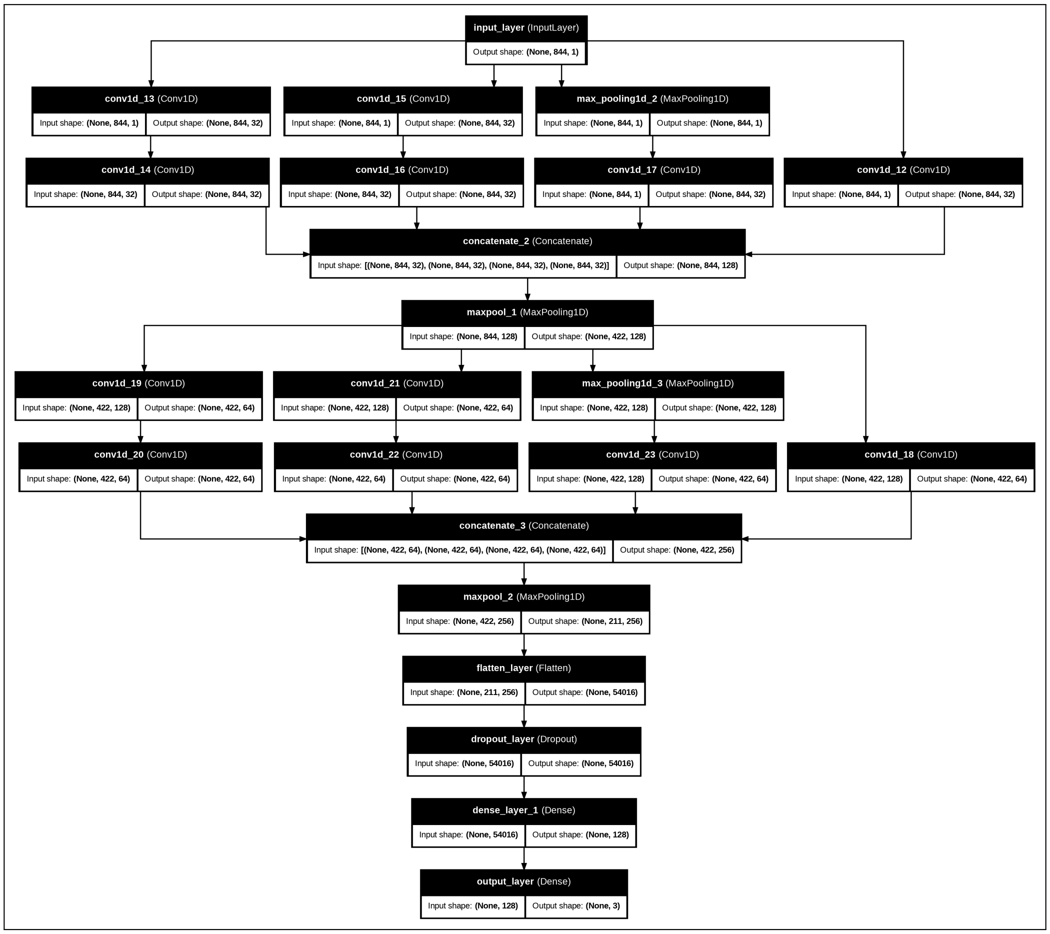

input_layer (InputLayer)
Output shape: (None, 844, 1)
conv1d_13 (Conv1D)
Input shape: (None, 844, 1) | Output shape: (None, 844, 32)
conv1d_15 (Conv1D)
Input shape: (None, 844, 1) | Output shape: (None, 844, 32)
max_pooling1d_2 (MaxPooling1D)
Input shape: (None, 844, 1) | Output shape: (None, 844, 1)
conv1d_14 (Conv1D)
Input shape: (None, 844, 32) | Output shape: (None, 844, 32)
conv1d_16 (Conv1D)
Input shape: (None, 844, 32) | Output shape: (None, 844, 32)
conv1d_17 (Conv1D)
Input shape: (None, 844, 1) | Output shape: (None, 844, 32)
conv1d_12 (Conv1D)
Input shape: (None, 844, 1) | Output shape: (None, 844, 32)
concatenate_2 (Concatenate)
Input shape: [(None, 844, 32), (None, 844, 32), (None, 844, 32), (None, 844, 32)] | Output shape: (None, 844, 128)
maxpool_1 (MaxPooling1D)
Input shape: (None, 844, 128) | Output shape: (None, 422, 128)
conv1d_19 (Conv1D)
Input shape: (None, 422, 128) | Output shape: (None, 422, 64)
conv1d_21 (Conv1D)
Input shape: (None, 422, 128) | Output shape: (None, 422, 64)
max_pooling1d_3 (MaxPooling1D)
Input shape: (None, 422, 128) | Output shape: (None, 422, 128)
conv1d_20 (Conv1D)
Input shape: (None, 422, 64) | Output shape: (None, 422, 64)
conv1d_22 (Conv1D)
Input shape: (None, 422, 64) | Output shape: (None, 422, 64)
conv1d_23 (Conv1D)
Input shape: (None, 422, 128) | Output shape: (None, 422, 64)
conv1d_18 (Conv1D)
Input shape: (None, 422, 128) | Output shape: (None, 422, 64)
concatenate_3 (Concatenate)
Input shape: [(None, 422, 64), (None, 422, 64), (None, 422, 64), (None, 422, 64)] | Output shape: (None, 422, 256)
maxpool_2 (MaxPooling1D)
Input shape: (None, 422, 256) | Output shape: (None, 211, 256)
flatten_layer (Flatten)
Input shape: (None, 211, 256) | Output shape: (None, 54016)
dropout_layer (Dropout)
Input shape: (None, 54016) | Output shape: (None, 54016)
dense_layer_1 (Dense)
Input shape: (None, 54016) | Output shape: (None, 128)
output_layer (Dense)
Input shape: (None, 128) | Output shape: (None, 3)

Fig. 4: Inception model

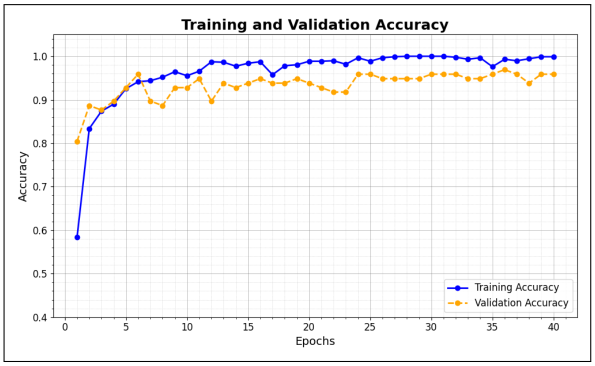


Fig. 5: Accuracy evolution over epochs

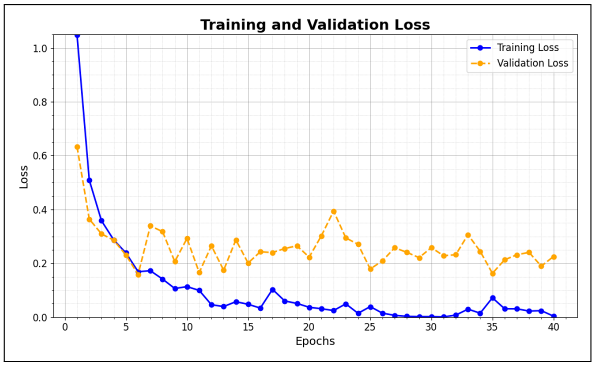


Fig. 6: Loss evolution over epochs

data alongside tactile-enhanced audiovisual content. By incorporating classifier-dependent feature selection, the study explored the synergy between audio-visual signals and neural responses, paving the way for novel applications in neurocomputing and immersive content analysis. Teng et al. [20] introduced a Transformer-based fusion model for depression detection. Their approach utilized intra- and inter-emotion constraints to enhance the fusion of multi-emotional audio-visual features. The inclusion of homogeneous and diverse constraints represents a unique contribution, particularly in detecting nuanced emotional states relevant to mental health applications. Sun et al. [21] developed HiCMAE, a hierarchical contrastive masked autoencoder for self-supervised audio-visual emotion recognition. Their innovative approach to leveraging self-supervised learning addressed challenges in labeled data scarcity, showing promise for robust emotion recognition in low-resource settings. Ying et al. [22] presented a multimodal driver emotion recognition algorithm designed for the Internet of Vehicles platform. By analyzing both audio and video signals, their work emphasized the

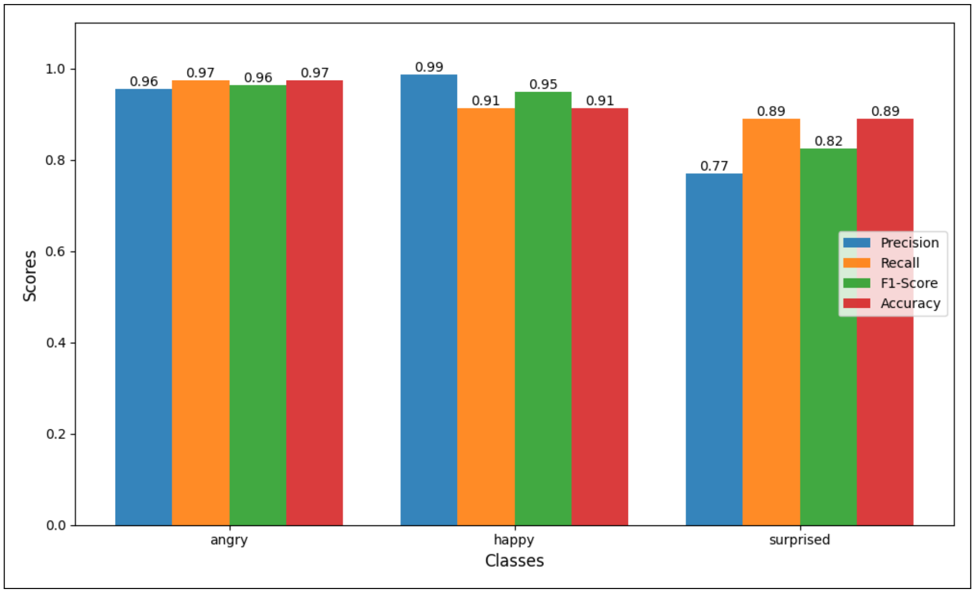


Fig. 7: Model performance

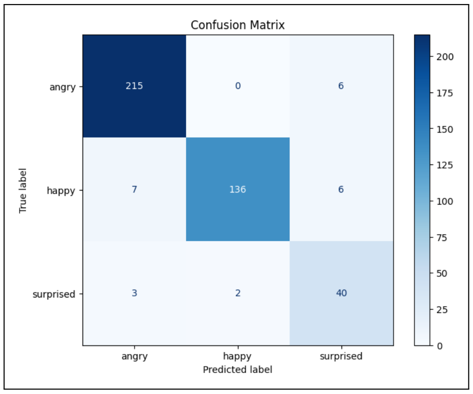


Fig. 8: Confusion matrix

importance of emotion-aware systems for improving driver safety and enhancing user experience in connected vehicle environments. Das et al. [23] proposed AVaTER, a cross-modal attention framework for fusing audio, visual, and textual modalities in emotion recognition. This study demonstrated how aligning information across modalities using attention mechanisms could enhance recognition performance, particularly in challenging multimodal datasets. Leem et al. [24] addressed the problem of noisy speech in emotion recognition. Their selective acoustic feature enhancement technique, tailored for speech emotion recognition, proved effective in mitigating noise and improving performance. This work is particularly valuable for real-world applications where audio quality is inconsistent. El Haj [25] extended the latent block model to emotion recognition in audio signals, offering a probabilistic approach to clustering emotional patterns. This methodology provided a statistical perspective to audio-based emotion recognition, highlighting its potential for interpretability and theoretical robustness. Meng et al. [26] introduced a masked graph learning method with recurrent alignment for multimodal emotion recognition in conversations. By aligning multimodal representations over temporal sequences, their framework effectively captured conversational context and emotion dynamics, offering a robust approach for dialogue-based applications. Wu et al. [27] proposed an audio multi-view spoofing detection framework that leverages audio-text-emotion correlations. Their work demonstrated how emotion analysis could strengthen spoofing detection in audio data, showcasing its potential in enhancing security and authentication systems. These studies illustrate the diverse methodologies and applications of emotion recognition, ranging from driver safety and mental health to audio spoofing detection and noisy environments. While models like Transformers, LSTMs, and autoencoders dominate the technical landscape, their success underscores the critical role of innovative data fusion techniques and noise-resilient architectures in advancing the field. Future research should focus on real-world deployments, handling data variability, and improving the interpretability of emotion recognition systems. Our present work aligns closely with the advancements described in these recent studies. Like these works, our approach leverages a deep learning architecture to tackle the challenges of emotion recognition, particularly in audio data. By using the Inception model, known for its ability to efficiently capture multi-scale features and subtle patterns, our framework ensures robust performance even in complex and noisy forensic audio environments. In future work, we aim to expand our Inception-based model by integrating multimodal data, such as video, textual transcripts, or physiological signals, to improve emotion recognition accuracy in complex forensic cases where audio alone may not suffice. Developing a real-time version of the model could enable on-the-fly emotion detection for applications like live interrogation analysis or emergency response systems. Additionally, we plan to adapt the model for diverse languages and cultural contexts by incorporating more varied datasets, ensuring broader applicability in global forensic scenarios.

VI. Conclusion

In this work, we proposed an Inception-based deep learning model for Arabic audio emotion recognition, designed specifically to address the unique challenges of forensic applications. Leveraging the Inception architecture's ability to efficiently extract multi-scale features, our model demonstrated robust performance in identifying subtle emotional cues from audio data, even in complex and variable conditions. The evaluation on a diverse dataset highlighted its accuracy, adaptability, and potential to handle real-world forensic scenarios where precision is critical. Our findings reinforce the importance of advanced neural architectures in bridging the gap between emotion recognition technology and forensic science. By providing a reliable and efficient tool for analyzing emotional states, this model can significantly contribute to areas such as interrogations, threat assessments, and credibility evaluations. While the current approach focuses on audio data, future research could explore integrating multimodal inputs, improving explainability, and enhancing resilience to noisy environments to further its applicability. This work underscores the potential of deep learning in transforming forensic investigations and sets the stage for continued advancements in the field of audio emotion recognition.

References

[1] J. Fattahi, "Machine Learning and Deep Learning Techniques used in Cybersecurity and Digital Forensics: a Review," *arXiv e-prints*, p. arXiv:2501.03250, Dec. 2024. https://ui.adsabs.harvard.edu/abs/2025arXiv250103250F.

[2] S. Qureshi, J. He, S. Tunio, N. Zhu, A. Nazir, A. Wajahat, F. Ullah, and A. Wadud, "Systematic review of deep learning solutions for malware detection and forensic analysis in iot," *J. King Saud Univ. Comput. Inf. Sci.*, vol. 36, no. 8, p. 102164, 2024.

[3] J. Fattahi, M. Mejri, and E. Pricop, *The Theory of Witness-Functions*, pp. 1–19. Cham: Springer International Publishing, 2016.

[4] J. Fattahi, M. Mejri, and M. Ziadia, "Extreme gradient boosting for cyberpropaganda detection," in *New Trends in Intelligent Software Methodologies, Tools and Techniques - Proceedings of the 20th International Conference on New Trends in Intelligent Software Methodologies, Tools and Techniques, SoMeT 202, Cancun, Mexico, 21-23 September, 2021* (H. Fujita and H. Pérez-Meana, eds.), vol. 337 of *Frontiers in Artificial Intelligence and Applications*, pp. 99–112, IOS Press, 2021.

[5] J. Fattahi, B. E. Lakdher, M. Mejri, R. Ghayoula, E. Manai, and M. Ziadia, "Fingfor: a deep learning tool for biometric forensics," in *2024 10th International Conference on Control, Decision and Information Technologies (CoDIT)*, pp. 1667–1672, 2024.

[6] J. Fattahi, *Analyse des protocoles cryptographiques par les fonctions témoins*. PhD thesis, Université Laval, Canada, February 2016.

[7] J. Fattahi, M. Mejri, M. Ziadia, and R. Ghayoula, "Spamdl: A high performance deep learning spam detector using stanford global vectors and bidirectional long short-term memory neural networks," in *New Trends in Intelligent Software Methodologies, Tools and Techniques - Proceedings of the 21st International Conference on New Trends in Intelligent Software Methodologies, Tools and Techniques, SoMeT 2022, Kitakyushu, Japan, 20-22 September, 2022* (H. Fujita, Y. Watanobe, and T. Azumi, eds.), vol. 355 of *Frontiers in Artificial Intelligence and Applications*, pp. 143–162, IOS Press, 2022.

[8] H. F. T. Alsaadawi, B. Das, and R. Das, "A systematic review of trimodal affective computing approaches: Text, audio, and visual integration in emotion recognition and sentiment analysis," *Expert Syst. Appl.*, vol. 255, p. 124852, 2024.

[9] S. Sadok, *Audiovisual speech representation learning applied to emotion recognition. (Apprentissage de représentation de la parole audiovisuelle pour la reconnaissance des émotions)*. PhD thesis, CentraleSupélec, Châtenay-Malabry, France, 2024.

[10] Z. Zhao, L. Alzubaidi, J. Zhang, Y. Duan, and Y. Gu, "A comparison review of transfer learning and self-supervised learning: Definitions, applications, advantages and limitations," *Expert Syst. Appl.*, vol. 242, p. 122807, 2024.

[11] A. Bajaj and D. K. Vishwakarma, "A state-of-the-art review on adversarial machine learning in image classification," *Multim. Tools Appl.*, vol. 83, no. 3, pp. 9351–9416, 2024.

[12] S. Klaylat, "Arabic Natural Audio Dataset (ANAD)." https://www.kaggle.com/datasets/suso172/arabic-natural-audio-dataset/data. Licensed under CC BY-NC-SA 4.0. Last accessed: Jan 12, 2025.

[13] B. McFee, C. Raffel, D. Liang, D. P. W. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python," in *Proceedings of the 14th Python in Science Conference 2015 (SciPy 2015), Austin, Texas, July 6 - 12, 2015* (K. Huff and J. Bergstra, eds.), pp. 18–24, scipy.org, 2015.

[14] Y. Jadoul, B. Thompson, and B. de Boer, "Introducing parselmouth: A python interface to praat," *J. Phonetics*, vol. 71, pp. 1–15, 2018.

[15] M. Magdin, T. Sulka, J. Tomanová, and M. Vozár, "Voice analysis using PRAAT software and classification of user emotional state," *Int. J. Interact. Multim. Artif. Intell.*, vol. 5, no. 6, pp. 33–42, 2019.

[16] A. Torfi, "Speechpy - A library for speech processing and recognition," *J. Open Source Softw.*, vol. 3, no. 27, p. 749, 2018.

[17] C. T. Batista, A. L. Dias, and N. Neto, "Free resources for forced phonetic alignment in brazilian portuguese based on kaldi toolkit," *EURASIP J. Adv. Signal Process.*, vol. 2022, no. 1, p. 11, 2022.

[18] Z. Jin and W. Zai, "Audiovisual emotion recognition based on bi-layer LSTM and multi-head attention mechanism on RAVDESS dataset," *J. Supercomput.*, vol. 81, no. 1, p. 31, 2025.

[19] A. Raheel, "Emotion analysis and recognition in 3d space using classifier-dependent feature selection in response to tactile enhanced audio-visual content using EEG," *Comput. Biol. Medicine*, vol. 179, p. 108807, 2024.

[20] S. Teng, J. Liu, Y. Huang, S. Chai, T. Tateyama, X. Huang, L. Lin, and Y. Chen, "An intra- and inter-emotion transformer-based fusion model with homogeneous and diverse constraints using multi-emotional audiovisual features for depression detection," *IEICE Trans. Inf. Syst.*, vol. 107, no. 3, pp. 342–353, 2024.

[21] L. Sun, Z. Lian, B. Liu, and J. Tao, "Hicmae: Hierarchical contrastive masked autoencoder for self-supervised audio-visual emotion recognition," *Inf. Fusion*, vol. 108, p. 102382, 2024.

[22] N. Ying, Y. Jiang, C. Guo, D. Zhou, and J. Zhao, "A multimodal driver emotion recognition algorithm based on the audio and video signals in internet of vehicles platform," *IEEE Internet Things J.*, vol. 11, no. 22, pp. 35812–35824, 2024.

[23] A. Das, M. S. Sarma, M. M. Hoque, N. H. Siddique, and M. A. A. Dewan, "Avater: Fusing audio, visual, and textual modalities using cross-modal attention for emotion recognition," *Sensors*, vol. 24, no. 18, p. 5862, 2024.

[24] S. Leem, D. Fulford, J. Onnela, D. Gard, and C. Busso, "Selective acoustic feature enhancement for speech emotion recognition with noisy speech," *IEEE ACM Trans. Audio Speech Lang. Process.*, vol. 32, pp. 917–929, 2024.

[25] A. E. Haj, "Emotions recognition in audio signals using an extension of the latent block model," *Speech Commun.*, vol. 161, p. 103092, 2024.

[26] T. Meng, F. Zhang, Y. Shou, H. Shao, W. Ai, and K. Li, "Masked graph learning with recurrent alignment for multimodal emotion recognition in conversation," *IEEE ACM Trans. Audio Speech Lang. Process.*, vol. 32, pp. 4298–4312, 2024.

[27] J. Wu, Q. Yin, Z. Sheng, W. Lu, J. Huang, and B. Li, "Audio multi-view spoofing detection framework based on audio-text-emotion correlations," *IEEE Trans. Inf. Forensics Secur.*, vol. 19, pp. 7133–7146, 2024.