2023 26th International Conference on Computer and Information Technology (ICCIT)
13-15 December, Cox's Bazar, Bangladesh

Multimodal Emotion Recognition through Deep Fusion of Audio-Visual Data

Tamanna Sultana[1], Meskat Jahan[1], Md. Kamal Uddin[2], Yoshinori Kobayashi[3], and Mahmudul Hasan[1*] *SMIEEE*
[1]*Department of Computer Science and Engineering, Comilla University, Cumilla, Bangladesh*
[2]*Department of Computer Science & Telecommunication Engineering, NSTU, Noakhali, Bangladesh*
[3]*Interactive Systems Lab., Saitama University, Saitama, Japan*
* mhasanraju@gmail.com

2023 26th International Conference on Computer and Information Technology (ICCIT) | 979-8-3503-5901-5/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICCIT60459.2023.10441424

***Abstract*—The field of emotion recognition in artificial intelligence focuses on enabling machines to comprehend and react to the range of emotions experienced by humans. This paper presents a novel approach that integrates the Convolution Neural Network (CNN) with audio and visual modalities. The study employs the RAVDESS database as a resource to train two distinct models for the analysis of both video and audio data. When it comes to audio pre-processing, advanced signal-processing techniques are applied to extract relevant elements and capture basic acoustic characteristics. A one-dimensional Convolutional Neural Network (CNN) architecture receives the audio data as input, enabling the model to learn complicated patterns and representations from the audio domain. In the context of video pre-processing, sophisticated algorithms are employed to extract essential facial characteristics. In order to capture the changing periods of facial expressions, the video frames are analyzed using a three-dimensional CNN framework following that they have been compressed and converted to grayscale. The fusion technique involves concatenating and extending the outputs of the audio and visual models. The fused features are subsequently sent into a softmax layer, which facilitates the development of a resilient emotion identification system.**

***Index Terms*—Emotion recognition, Multi-modal fusion, Convolution Neural Networks, Audio-visual recognition.**

I. INTRODUCTION

In an increasingly interconnected and technologically driven world, understanding and interpreting human emotions is of paramount importance. Future human environments are anticipated to incorporate a range of technologically advanced sensors that can help and anticipate essential actions in controlling their feelings in a vast and invisible way [1]. However, since the manifestations connected to human emotions vary greatly among people and civilizations, recognizing emotions from the human environment is a difficult subject [2]. Scholars continue to be relentlessly working to make it possible for models to understand and recognize human-like emotions [3].

Different modalities, including speech, facial expression, and text analysis, can be used to extract emotions. When it comes to immediate response, audio-only methods may encounter problems like noisy data addition or improper coefficient collection, while visual-only methods may encounter problems like occlusions or facial movements that these methods might misinterpret, like yawning. However, both visual and audio information combined with one other helps to get around these challenges. Here, the significance of the audio-visual fusion concept is highlighted. While a number of techniques have been put out in recent years to identify emotions from speech or visual cues, less focus has been placed on fusing these two in Emotion Recognition (ER) [4], [5]. This paper's goal is to extract an emotion from an audio as well as a video clip.

Three architectural frameworks for emotion recognition are proposed in this work: audio model, video model, and fusion model. We use the output of the suggested 1D CNN model as a feature vector for speech-based recognition. In the face-based technique, we capture expressions from clips of video and utilize a 3-dimensional network to determine emotions. We combine the results of these two approaches to develop a multi-modal emotion identification model that takes advantage of both speech and facial cues. Our suggested model successfully extracts emotional information from audio-visual input, achieving an accuracy of 66.09% when tested against the RAVDESS database [6].

II. RELATED WORK

Since there is no explicit mapping between an individual's sentiment state and both visual & audio components in many recognition of emotions methods, effective emotion information extraction from audio and visual data is challenging [7]. Numerous scholars have developed many approaches and frameworks in recent years to address the problems with effective emotion recognition.

The researchers at [8] present a feature fusion technique, which successfully performs intermediate integration regarding obtained characteristics, and outperforms fusion-related methods, while offering cutting-edge performance on two large databases. According to the authors of article [9], two separate types of fusion can be used by neural networks to merge several visible modes and an audio channel in order to identify the active speaker. They went with a 3D CNN architecture since it is efficient at simultaneously encoding movement and appearance. In [10], a multi-modal facial expression identification approach is presented that uses both face images and audio data to distinguish between ambiguous facial emotions. The authors explicitly provide a Modal Fusion Module, where image and audio components are extracted from Swin Transformer, to fuse audio-visual data. They also use dynamic data re-sampling to address the imbalance problem in the database. According to the authors of [3], continuous speech modeling, which imitates human comprehension, enhances the ability

to gather emotional data when compared to quantized methods. They highlight its importance by proposing a unique search space that combines speech information with textual semantics. Their MFAS framework often outperforms current speech emotion recognition technologies and mimics human emotional comprehension. The authors of article [11] give a variety of thorough methods for categorizing human emotions, such as VGG-LSTM based on Faces and Acoustic SVM Classifier.

In conclusion, recent research has shown that combining several data modalities and fusion techniques can significantly increase the efficacy of emotion detection systems. The use of Convolutional Neural Networks (CNN) for categorizing has also been shown to improve system accuracy. By developing a potent fusion and CNN-based emotion recognition system, the proposed work aims to advance this field.

III. RESEARCH METHODOLOGY

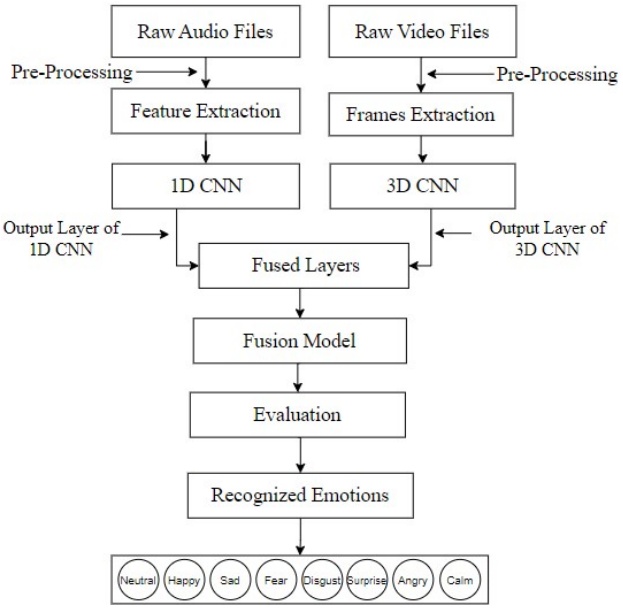

Fig. 1. System Architecture

Previous systems were found to be inaccurately deficient based on a recent investigation in a comparable field. As a result, we suggest a method for recognizing emotions that will function well with the data and its pertinent aspects. Fig. 1 depicts the overall block diagram of the proposed system. Audio & video are the two distinct forms of information the system accepts. Speech and sight signals have been assessed independently and examined prior to categorization, and they are subsequently combined. Before fusion, each of these modalities goes through two essential steps: initial processing and deep networks.

The input, which consists of audio and video samples, is pre-processed. Pre-processing involves a number of methods, such as face detection, photo trimming, adjusting the size and retrieval from the visual feed, division and overlapping, multiplying frames, and many more for the audio input. The appropriate CNN's are then implemented. The second last output layer from each CNN is then concatenated and fed into a fusion layer. After evaluating the corresponding fusion features, the desired outputs are achieved.

A. Pre-processing

1) Input Audio Pre-processing: Data pre-processing is necessary to guarantee the correctness and efficiency of the model. It is done before the specific characteristics of the speech examples are retrieved. While the speech was being recorded, unwanted information like noise and environmental changes were recorded. Hence, the removal of this unnecessary information is a must. The audio pre-processing in our work that is displayed in Fig. 2, which incorporates a series of tasks-

- Using the specified mapping, an emotion label attached to every audio file is derived from the filename.
- Required libraries are used to separate characteristics of sound from an audio stream.
- After processing all the audio files, the next step involves making the audio features consistent in length.
- Padding & trimming are applied to match the desired length.
- The Final pre-processed audio features are now to be used as input to a machine-learning model for emotion recognition or other tasks.

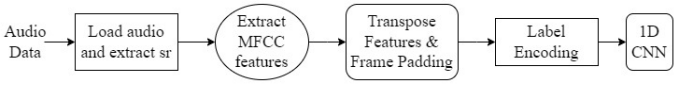

Fig. 2. Audio pre-processing block diagram

2) Input Video Pre-processing: We present a powerful pipeline for pre-processing videos that is designed specifically for the Ravdess dataset. The video file directory is used to calculate the frame rate. After that, related frames are created for each video file with a duration of 3 seconds. For consistency, each frame is enlarged to a standard (80x70) size. While padding or truncating guarantees a set frame count (30 frames) for consistency, an exclusion function removes unnecessary frames from the beginning and conclusion of the film. Using a mapping dictionary, emotion labels are extracted from filenames. For following emotion detection tasks and scholarly investigation into video-based emotion recognition, this sophisticated pre-processing pipeline establishes a solid foundation. Fig. 3 depicts the overall pre-processing steps.

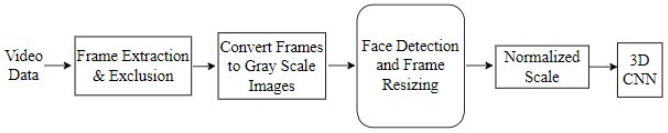

Fig. 3. Video pre-processing block diagram

Frames that are extracted from the video file are in grayscale with a resolution of (80x70) pixels. 30 frames are extracted

from each 3-second video file. The following figure displays some retrieved frame examples.

Fig. 4. Extracted Frames [6]

B. CNN Architecture

Convolution and non-linearity tasks are used by CNN as a superb pattern acquisition and retrieval approach to understand the characteristics of the neighborhood and spatial surfaces [12]. The proposed framework employs a specific kind of framework for both modalities.

1) Audio CNN Architecture: 1D-CNN is suggested for audio input data in order to create a foundation for emotion recognition. In this architecture, a combination of convolution & max-pooling layers are constructed and the final layer is established by a fully associated neural structure with two dense layers. A softmax work is provided for the corresponding layer yield overall. Two fully connected dense layers with 256 and 64 units, both using ReLU activation, capture complex patterns. Emotion probabilities for 8 classes (such as neutral, happy, and sad) are produced by the last dense layer. A softmax layer is applied after the final yield layer to ensure fair distribution of the yield values. Fig. 5 depicts the proposed audio CNN architecture.

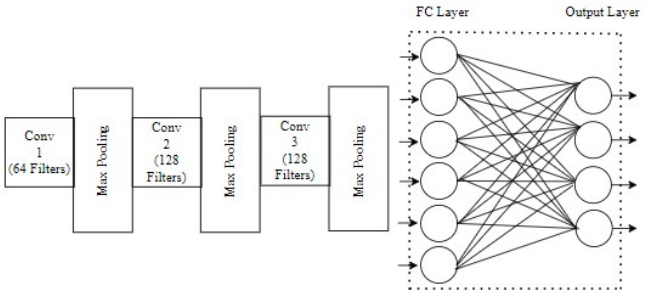

Fig. 5. Audio CNN Architectural Diagram

2) Video CNN Architecture: The creation of an emotion classification paradigm for the visual stream has been suggested using a 3-dimensional convolution neural network. The combination of two max-pooling layers with four convolution layers established the architectural framework. the final layer is established by a neural structure with two dense layers. A softmax work is provided for the corresponding layer yield overall. There are 16 channels with kernel size (3x3x3) in the first layer, 16 channels with the kernel (3x3x3) in the second layer, and 32 channels in the last two layers. Complex patterns are captured by a fully interconnected layer with 64 units and ReLU activation. The final layer is used for showing emotion probabilities for 8 classes (e.g., neutral, happy, sad). Fig. 6 depicts the architectural representation of the proposed video CNN architecture. Emotions can be recognized from this model, but for computational purpose we use only the second last layer for fusion input.

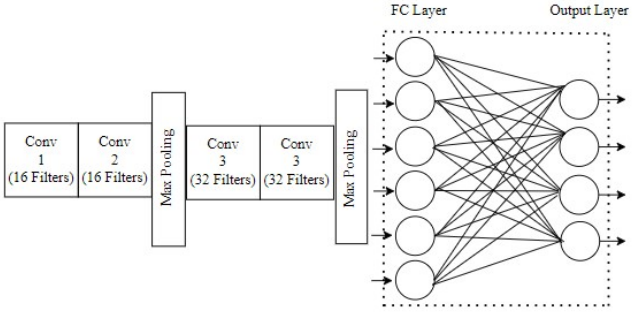

Fig. 6. Video CNN Architectural Diagram

3) Fusion Model: Models built using deep learning are excellent for object recognition systems because of their broad adoption, particularly CNN's [12]. The fusion technique enables more robust and accurate emotion analysis based on various input data sources by merging the data from both video and audio modalities to improve emotion detection performance. As a result, we suggested a fusion model that is also built on the CNN architecture. The second-to-last layers from the trained audio and video models are combined in the proposed fusion strategy. The output is concatenated and then a dense layer with 128 units and a ReLU activation function is used, as suggested by [13]. Another dense layer has been used to generalize the model's output. The final output of the fusion model represents the probabilities for 8 potential classes (for example, emotions) in the multi-class classification problem. Fig. 7 depicts the architectural framework for the fusion approach.

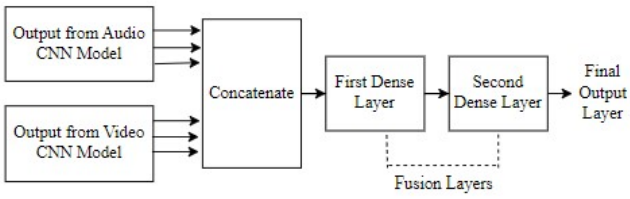

Fig. 7. Architectural Diagram of Fusion Model

IV. Experimental Result

A. Database

RAVDESS dataset [6] has been chosen for this research. This is an established multidisciplinary collection of moving speech and music that has been verified. 7356 files totaling 24.8 GB make up the database. The collection contains recordings of 24 professional actors (12 men and 12 women) vocalizing two lexically comparable sentences in a neutral

North American accent. Three modalities are accessible for each scenario: audio-only, video-only, and audio-video. However, we have used 2880 files (1440 video-only files and 1440 audio-only files) for our research.
Each file within RAVDESS possesses a unique filename. Hyphens are used to separate the seven two-digit numerical identification numbers that make up the filename (e.g., 02-01-06-01-02-01-12.mp4). Every two-digit number represents the level of a unique experimental factor. These are the identifiers: mp4 or.wav, modality, channel, emotion, intensity, statement, repetition, and actor. The case distribution for each emotion is shown in Table I.

TABLE I
CASE DISTRIBUTION UNDER EMOTIONS

Emotion	No of Case
Fear	192
Angry	192
Sad	192
Surprise	192
Calm	192
Happy	192
Disgust	192
Neutral	96
Total	1440

B. Experimental Setup

The virtual machine was used for every simulation and operations, comprising network testing and training, with the system configurations indicated in Table II. The CNN models with the following parameters were used: validation split = 0.4 and 0.2, learning rate = 0.001, and batch size = 32. For three distinct models, epoch counts of 25, 20, and 12 were employed.

TABLE II
SYSTEM CONFIGURATIONS

Model Name	Intel(R) Xeon(R) CPU @ 2.30GHz
CPU Cores	16
CPU MHz	2300.000
RAM	30 GB
GPU	16 GB

C. Testing Accuracy

Table III summarizes the testing set accuracy obtained among different architectures. We have proposed three different architectures: audio, video, and fusion architecture.
For the audio paradigm, low-level features are extracted from the speech signal, and a 1D CNN is used. For each spectrogram, a frame size of (180x20) has been considered. In the training phase, each spectrogram is considered a sample. For testing, all the spectrograms of a speech signal are fed into the proposed model. The average accuracy achieved with this method is 64.09%.
For the video paradigm, the frames are extracted from each video. Each frame is an image with dimensions of (80x70). Sufficient samples are produced for training. In the test phase, the key faces for a sample video are fed into the model, and then the final result is obtained using the last softmax layer. The average accuracy achieved with this method is 59.07%. The experiments show that the speech-based emotion recognizer achieves more accurate results in comparison with the facial-based emotion recognizer.
Ultimately, the multi-modal framework that is suggested and shown in Figure 1 is justified. This approach yields an average accuracy of 66.90%. It is evident that taking voice and picture information into account has improved the ability to identify emotions. Taking this into account, Table III presents a comparison of several techniques.

TABLE III
TESTING ACCURACY OVER AUDIO, VIDEO & FUSION MODEL

Architecture	Accuracy(%)
1D CNN(Audio)	64.09
3D CNN(Video)	59.07
Fusion	66.90

D. Loss & Accuracy Curve on Fusion Model

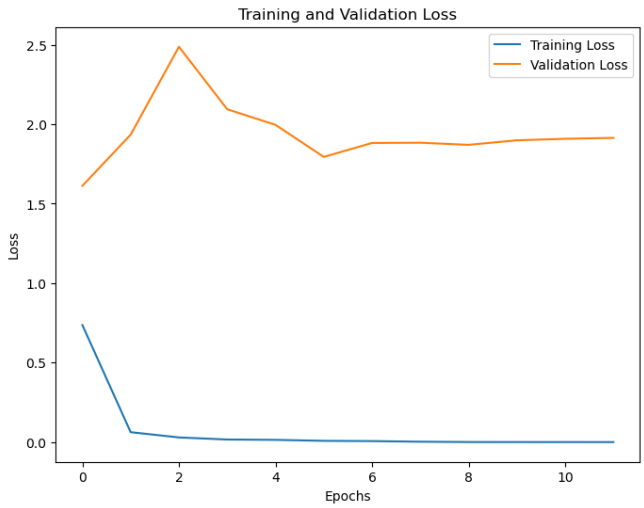

Fig. 8. Training & Validation loss curve

In Fig. 8, we observe the model's exceptional performance on the training set, reflected in the small training loss. However, there is an opportunity for improvement as the validation loss appears larger, indicating a potential area for optimization in generalization. It is noteworthy that the combination of activation layers from diverse models has significantly enhanced the model's ability to adapt to the training data.
Examining the accuracy comparison illustrated in Fig. 9, it is evident that the training accuracy surpasses the validation accuracy. It has also been displayed how many epochs were used for the fusion approach. As can be noticed, our approach performs better when using the training dataset. The results

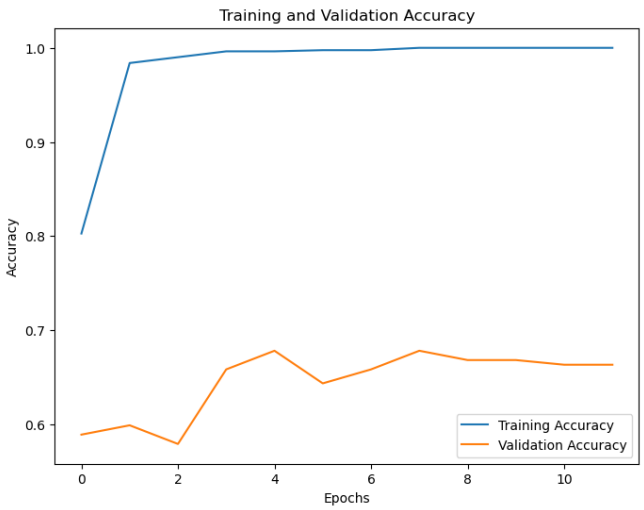

Fig. 9. Testing & validation accuracy

of this investigation demonstrate that it is possible to conduct fruitful research in multi-modal emotion recognition.

E. Confusion Matrix on Fusion Model

The normalized confusion matrix is shown in Fig. 10, with the x-direction showing expected labels and the y-direction showing genuine labels. According to the labels on the audio model's evolution, this matrix displays the actual prediction made by our model. We discovered that the label **calm**, which predicts 26 cases, is the model's highest prediction, while the label **neutral**, which only predicts 7 cases, is its lowest.

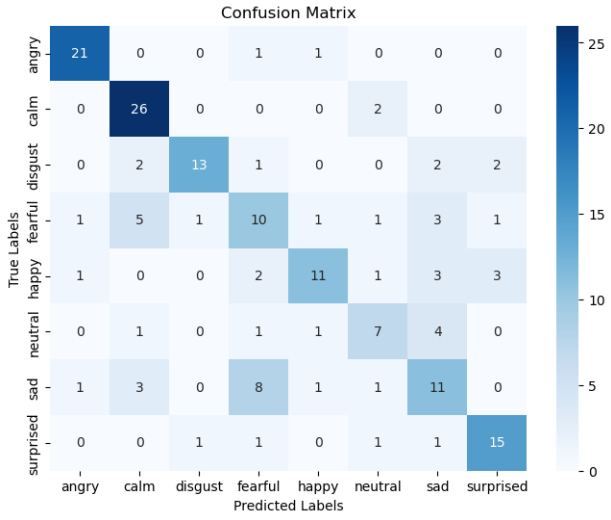

Fig. 10. Training performance of the fusion model

The overall performance could have been better if we could use a dedicated machine with the specifications described instead of a shared virtual machine with limited resource usage. With 2880 audio and video files in our huge dataset, the RAM and CPU memory were not up to the mark. The VM was only allowed to use two cores from the actual CPU, which had 16 cores in it. Moreover, only (80x70) pixel frames are taken from video files. Additionally, we could improve the overall performance of our method if we could extract high-resolution frames like (224x224).

V. Conclusion

In this work, we used a fusion model that combines auditory and visual data to do a thorough analysis of emotion identification. The proposed fusion model incorporates features from the Conv1D and Conv3D architectures, which were specifically designed for the processing of audio and video data, respectively. The proposed approach leverages the limitations of uni-modal methodologies by including many modalities, resulting in improved accuracy and robustness in the task of emotion recognition. Our model has a high level of proficiency in recognizing the emotion of calmness among the various emotions with a maximum accuracy of 66.90%.
In our forthcoming research, we intend to assess the performance of the model on more extensive and varied emotion identification datasets, such as AffectNet or SAVEE. This endeavor aims to gain a deeper understanding of the model's practical utility in real-world scenarios.

References

[1] Juan D. S. Ortega, Mohammed Senoussaoui, Eric Granger, Marco Pedersoli Patrick Cardinal and Alessandro L. Koerich, "Multimodal Fusion with Deep Neural Networks for Audio-Video Emotion Recognition," arXiv:1907.03196v1 [cs.CV] 6 Jul 2019.
[2] Gnana Rajasekar, Wheidima Melo, Nasib Ullah, Muhammad Haseeb Aslam, Osama Zeeshan, Denorme, Marco Pedersoli, Alessandro Koerich, Patrick Cardinal, and Eric Granger. "A joint cross-attention model for audio-visual fusion in dimensional emotion recognition". 03 2022.
[3] Fulin Zhang, Zheng Lian, Yingying Gao, Shilei Zhang, Hai Yang, "MFAS: Emotion recognition through multiple perspectives fusion architecture search emulating human cognition," China Mobile Research Institute,Institute of Automation, Chinese Academy of Sciences, 2023.
[4] Khadijeh Aghajani. "Audio-visual emotion recognition based on a deep convolutional neural network," Journal of Artificial Intelligence and Data Mining (JAIDM), 2022.
[5] Syrine Haddad, Olfa Daassi, Safya Belghith. "Emotion Recognition from Audio-Visual Information based on Convolutional Neural Network", 2023 International Conference on Control, Automation and Diagnosis (ICCAD), 2023.
[6] Livingstone SR, Russo FA (2018) 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 13(5): e0196391.
[7] Min Chen Jincai Chen Ping Lu Andrej Ko˘sir Yaxiong Ma, Yixue Hao. "Audio-Visual Emotion Fusion (AVEF): A deep efficient weighted approach," Information Fusion Volume 46, March 2019, pages 184–192, 2019.
[8] Zhong, D. Schneider, M. Voit, R. Stiefelhagen and J. Beyerer, "Anticipative Feature Fusion Transformer for Multi-Modal Action Anticipation," 2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 2023, pp. 6057-6066, doi: 10.1109/WACV56688.2023.00601.
[9] Pibre, L., Madrigal, F., Equoy, C. et al. Audio-video fusion strategies for active speaker detection in meetings. Multimed Tools Appl 82, 13667–13688 (2023). https://doi.org/10.1007/s11042-022-13746-7
[10] Kim Junhwa, Namho Kim, and Chee Won. Multi-modal facial expression recognition with transformer-based fusion networks and dynamic sampling. 03 2023.
[11] Xin Guo, Luisa Polania, and Kenneth Barner. Audio-video emotion recognition in the wild using deep hybrid networks. 02 2020.
[12] Amashita, R., Nishio, M., Do, R.K.G. et al. Convolutional neural networks: an overview and application in radiology. Insights Imaging 9, 611–629 (2018). https://doi.org/10.1007/s13244-018-0639-9.
[13] Adamantios Ntakaris, Giorgio Mirone, Juho Kanniainen, Moncef Gabbouj, Alexandros Iosifidis. "Feature Engineering for Mid-Price Prediction With Deep Learning", IEEE Access, 2019.