Audio-Visual Emotion Recognition Using Multilevel Fusion

Muhammad Shoaib[*], Sana Ul Haq[†], Muhammad Saeed Shah[‡], Imtiaz Rasool[§], Mohammad Omer Farooq[**]

Abstract
In the affective computing domain, many researchers have worked on automatic human emotion recognition in recent years. Unimodal techniques, i.e., audio, visual or physiological signals, have been used in most emotion recognition research. The research indicates that one modality can trump the other when it comes to classification accuracy. Some emotions may have better classification accuracy in one modality, while others may be easily separated in the other. In the proposed research, emotion recognition is performed using both unimodal and bimodal techniques. Experiments were performed using six emotions of an audio-visual interactive emotional dyadic motion capture (IEMOCAP) database. The classification was performed using three different feature selection methods and seven various classification techniques. The recognition accuracy of 64.54% was obtained for the audio modality, and 96.77% for the visual modality using rotation forest classifier. For the bimodal approach, the best accuracy of 96.04% was obtained for feature-level fusion using the rotation forest classifier. The decision-level fusion resulted in the best performance of 97.60% for the product rule, while obtained an accuracy of 97.51% for the sum rule. The bimodal approach provided better results in comparison to unimodal approach, and the decision-level fusion provided better results.

***Keywords*:** Emotion Recognition; Decision-Level Fusion; Sum Rule; Product Rule; Classification

Introduction

Nowadays, multimodal emotion recognition is getting more attention from scientists and researchers due to market demands for intelligent technologies and a wide range of applications (Khare et al., 2024). Effective emotion recognition in the field of robotics technology can provide individuals with a friendlier interactive environment. As a result, automatic multimodal emotion recognition has attracted much devotion in the real-world scenarios (Zhou et al., 2021). Emotion

[*] Department of Electronics, University of Peshawar, Peshawar 25120, Pakistan, mshoaib@uop.edu.pk
[†] Department of Electronics, University of Peshawar, Peshawar 25120, Pakistan, sanaulhaq@uop.edu.pk
[‡] Department of Electronics, University of Peshawar, Peshawar 25120, Pakistan, saeedshah@uop.edu.pk
[§]Corresponding Author: Department of Electronics, University of Peshawar, Peshawar 25120, Pakistan, imtiazrasoolkhan@uop.edu.pk
[**] Department of Electronics, University of Peshawar, Peshawar 25120, Pakistan, omer908@uop.edu.pk

recognition technology has numerous applications in various fields of life, such as medicine (Yuvaraj et al., 2014), service robots, customer care, advertising, and entertainment (Khalil et al., 2019). Machines must be able to perceive and mimic verbal and facial expressions for human-computer interaction to be more efficient and realistic (Samadiani et al., 2019). To enhance the machine intelligence, researchers have focused on both the audio and visual emotion recognition.

Audio Emotion Recognition

Verbal communication is the direct channel of communication in human interaction. From the properties of audio signals, such as voice quality, prosodic expression in pitch, rhythm, and energy contours, people can recognize different emotions. Several types of acoustic features, such as continuous, qualitative, and spectral features, have been utilized to recognize audio emotions (Zhao et al., 2019). The prosodic features have been observed to be very useful in emotion recognition (Ma et al., 2019). Araño et al. (2021) combined the spectrogram image features and Mel Frequency Cepstral Coefficients (MFCCs) for audio emotion recognition. The MFCCs features with Long Short-Term Memory (LSTM) network outperformed the Support Vector Machine (SVM) classifier. Mannepalli et al. (2022) used multiples support vector neural network classifier for audio emotion classification to achieve improved results. Alluhaidan et al. (2023) proposed Convolutional Neural Network (CNN) which provided improved classification results by combining the MFCCs and time-domain features (MFCCT). Mohan et al. (2023) used a combination of 2D CNN and eXtreme Grading Boosting (XG-Boost) with MFCC features to achieve better classification performance. Bhanusree et al. (2023) used a CNN to extract features, while the Random Forest was used for classification. The proposed technique obtained higher classification accuracy on IEMOCAP and RAVDESS datasets.

Visual Emotion Recognition

The facial expressions effectively convey human emotional information. The accuracy of audio emotion recognition does not meet the market standards, hence many researchers focused on visual emotion recognition to improve the classification accuracy. It describes all emotional states that are expressed through variations in the muscles of the face, eyes, and mouth. The most noticeable of these are the muscles that surround the mouth and eyes (Ekman, 1993). Tzirakis et al. (2017) used speaker face normalization, principal component analysis, and principal feature analysis to obtain compact facial representation in order to extract visual features. Visemes were measured for the facial expressions of

various emotion classes. The accuracy of 75% was obtained for happiness, 50% for anger, 60% for sadness, and 35% for neutrality. The viseme information increased the overall classification performance. Ekman & Friesen (1978) created the facial action coding system (FACS) to describe facial expressions. All possible facial expressions in FACS are broken down in action units (AUs). Zhang et al. (2017) extracted 15 visual features using a general-purpose tracking algorithm. The visual information was obtained by recognizing the positions of eyebrows, cheeks' lift, and opening of the mouth. Bota et al. (2020) used an optical flow algorithm to extract visual features by recognizing the edge moment of brows, lips, and mouth corners. Haq et al. (2015) extracted 290 visual features related to face marker's positions and angles.

Audio-Visual Emotion Recognition

In recent years, researchers have focused on audio-visual emotion recognition. The two modalities are fused at feature, decision, and model levels. The feature-level fusion has been performed in many studies (Wang et al., 2012). But the feature-level fusion failed to describe the complex interactions between the two modalities such as disparities in time scales and metric levels (Zhang et al., 2017). Most emotion recognition challenges involve decision-level fusion (Thiam et al., 2020). It is normally performed by combining the individual categorization scores, due to which it cannot capture the mutual association between distinct modalities very well. In some studies, the audio and visual information were integrated using the hidden Markov models (HMMs) (Zeng et al., 2008). In Zhao et al. (2009), the mouth was divided into several sub-regions for the extraction of Local Binary Patterns from Three Orthogonal Planes (LBP-TOP) features from every sub-region. The results were then combined to improve the classification accuracy. By simulating the human emotion recognition systems, multiple attention fusion networks were introduced by Wang et al. (2019).

Most researchers have used the unimodal approaches, i.e., audio, visual, or physiological signals, to study emotion recognition (Chen et al., 2018; Tan et al., 2021; Chalapathi et al., 2022). The overall classification accuracy of unimodal approaches is lower because the single modality carries less information about the emotional state of a person. For this reason, multimodal techniques have been proposed for emotion recognition to improve the classification performance (Nie et al., 2020; Avots et al., 2019). The proposed research concentrates on multimodal fusion for audio-visual emotion recognition employing both feature-level and decision-level fusion. The following sections present the methodology, experimental results, and conclusion.

Methodology

Firstly, the IEMOCAP database was acquired. It was followed by feature extraction, feature normalization, feature selection, and classification.

Emotional Database

Different databases have been used to design the emotion recognition systems. The audio-visual IEMOCAP database (Busso et al., 2008) is utilized for experiments in this research. The database was recorded from ten actors (five male and five female) in a mixed-gender paired form. The markers were used on the face, head, and both hands. The collected visual data were three-dimensional. The database contains five sessions, each of which contains dialogues and sentences. The entire dialogue is made up of improvised and scripted data. This database contains ten emotions that have been classified by three annotators. The visual data is available as a text file, while the audio data is available as a wav file. The IEMOCAP database was chosen due to its large size, and it covers the basic human emotions. In addition, it contains the face markers data for the extraction of visual features. Other audio-visual databases such as SAVEE (Haq & Jackson, 2011) have limited data, while the RAVDESS (Livingstone & Russo, 2018) does not contain the face marker data.

Feature Extraction

To extract features from both audio and visual data various software were used. The feature extraction module receives the input in the form of audio or visual signal, and outputs a feature vector. The OpenSMILE toolkit (Eyben et al., 2009) was used for the extraction of 6540 audio features, while the MATLAB software (The MathWorks Inc., 2024) was utilized to extract 494 visual features.

Feature Normalization

In a classification task, feature normalization is an essential part that limits the raw features data to a fixed range. For data scaling, Min-Max (Jain & Bhandare, 2011) and Z-Score (Jain et al., 2005) normalization approaches are normally used. In the proposed research, Min-Max normalization was applied through the Weka toolkit (Witten et al., 2016).

The Min-Max normalization in the range $[r_{min}\ r_{max}]$ is defined by the relation

$$\overline{k} = \frac{k - k_{min}}{k_{max} - k_{min}} \times (r_{max} - r_{min}) + r_{min} \quad (1)$$

where $\bar{k}$, k_{min}, and k_{max} are the normalized, minimum and maximum values of attribute k.

In the proposed research a range of [0, 1] was used, for which the equation (1) becomes

$$\overline{k} = \frac{k - k_{min}}{k_{max} - k_{min}} \quad (2)$$

Feature Selection

The procedure of feature selection selects the most important features from the extracted features. It decreases the computational complexity while improving the classification performance. This process retains the most valuable features while eliminating the unrelated features.

In the proposed approach, different kinds of attribute evaluators were used for feature selection, i.e., Correlation-based Feature Selection (CFS) (Hall, 1999), Info Gain (Azhagusundari & Thanamani, 2013), and Gain Ratio (Witten et al., 2016). The CFS technique selected the subsets of features using the best first and greedy stepwise search methods, while the Info Gain and Gain Ratio techniques ranked the individual features. Table 1 shows the numbers of selected features for audio and visual modalities. The Weka toolkit was used for feature selection.

Table 1: Audio and visual features selected using different techniques.

Attribute Evaluator	Search Method	Number of Selected Features		
		Audio	Visual	Audio-Visual
CFS	Best First	219	33	209
	Greedy Stepwise	220	36	214
Info Gain	Ranker	3000	494	3000
Gain Ratio	Ranker	3000	494	3000

Classification

The final step in emotion recognition is the classification of different emotions. In this research, the emotion recognition was performed using seven different classification techniques, i.e., Bayes Net (Liu et al., 2016), SVM (Chang & Lin, 2011), Meta Bagging (Büchlmann & Yu, 2002), Rotation Forest (Rodriguez et al., 2006) , Functional Trees (Gama, 2004), Random Forest (Breiman, 2001), and Random Trees (Witten et al., 2016).

Fusion at Feature and Decision Levels

The use of single modality (audio or visual) results in overall lower classification performance due to limited available information. To enhance the classification accuracy, the different modalities are fused at

different levels, i.e., feature, decision and model. The proposed research investigates the feature and decision levels fusion. In feature-level fusion, the audio and visual features were combined in a single feature vector. Afterwards, feature normalization, selection and classification were performed. In decision-level fusion, the classification outputs of the two modalities were combined using the sum and product rules.

Experimental Results

The classification results were obtained for six emotion classes, i.e., anger, excited, frustration, happy, neutral and sadness, of the IEMOCAP database using 10-fold cross validation.

Audio Emotion Classification

The audio emotion recognition was performed using three different feature selection methods. The best results for CFS with best first and greedy stepwise search methods, info gain, and gain ratio feature selection techniques are given in Table 2. The best result of 64.40% was obtained for the CFS with best first search method and rotation forest classifier, while the best accuracy of 64.54% was obtained for the CFS with greedy stepwise search method and rotation forest classifier. The meta bagging classifier obtained the best classification accuracy of 64.54% using the info gain method, and 64.15% using the gain ratio technique. For audio emotion recognition, the best accuracy of 64.54% was obtained for both the CFS (greedy stepwise search method) with rotation forest classifier, and info gain feature selection with meta bagging classifier.

Visual Emotion Classification

The visual emotion classification results are given in Table 3 for the CFS with best first and greedy stepwise search methods, info gain, and gain ratio feature selection techniques. For the CFS with best first search method the best classification accuracy of 96.40% was obtained using the rotation forest classifier, while the CFS with greedy stepwise search method resulted in the best result of 96.40% with random forest classifier. The info gain feature selection method resulted in the best accuracy of 96.77% with rotation forest classifier, while the gain ratio feature selection method obtained the best result of 96.70% with rotation forest classifier. The best visual emotion classification result of 96.77% was obtained for the info gain feature selection with rotation forest classifier.

Table 2: Audio emotion recognition results.

Attribute Evaluator	Search Method	Classifier	Classification Accuracy (%)
CFS	Best First	Rotation Forest	64.40
	Greedy Stepwise	Rotation Forest	64.54
Info Gain	Ranker	Meta Bagging	64.54
Gain Ratio	Ranker	Meta Bagging	64.15

Table 3: Visual emotion recognition results.

Attribute Evaluator	Search Method	Classifier	Classification Accuracy (%)
CFS	Best First	Rotation Forest	96.40
	Greedy Stepwise	Random Forest	96.40
Info Gain	Ranker	Rotation Forest	96.77
Gain Ratio	Ranker	Rotation Forest	96.70

Audio-Visual Emotion Classification

The feature and decision levels fusion were used for the bimodal emotion recognition. Table 4 shows the classification results obtained for feature-level fusion using CFS with best first and greedy stepwise search methods, info gain, and gain ratio feature selection techniques. The rotation forest classifier resulted in the best classification accuracy of 94.85% for CFS with best first search method, while it achieved the best result of 94.92% for CFS with greedy stepwise search approach. For the info gain and gain ratio feature selection approaches, the rotation forest classifier resulted in the best classification performance of 96.04% and 94.92%, respectively. The best bimodal emotion recognition result of 96.04% was achieved using the info gain feature selection with rotation forest classifier.

Table 4: Audio-visual emotion recognition results for feature-level fusion.

Attribute Evaluator	Search Method	Classifier	Classification Accuracy (%)
CFS	Best First	Rotation Forest	94.85
	Greedy Stepwise	Rotation Forest	94.92
Info Gain	Ranker	Rotation Forest	96.04
Gain Ratio	Ranker	Rotation Forest	94.92

The sum and product rules were used for the decision-level fusion. The results are given in Table 5. For the product rule, the classification accuracies of 96.40%, 68.47%, 96.80%, and 97.60% were obtained for CFS with best first and greedy stepwise search methods, gain ratio, and info gain feature selection techniques, respectively. For the sum rule, the

best accuracies of 96.43%, 96.60%, 97.51%, and 97.50% were obtained for the CFS with best first and greedy stepwise search methods, gain ratio, and info gain feature selection techniques, respectively.

The confusion matrix for the best classification accuracy obtained for the decision-level fusion using the product rule is given in Table 6. The best result was obtained for the frustration emotion, followed by sadness, and excited. The anger emotion was confused with frustration which resulted in lower accuracy for the anger emotion. The lowest performance was obtained for the happy emotion due to confusion with excited, neutral and sadness.

The best classification results obtained for the unimodal and bimodal scenarios are summarized in Table 7. These results indicate significantly better performance for the visual and bimodal scenario as compared to audio modality. The best performance of 64.54%, 96.77%, and 97.60% was obtained for the audio, visual, and audio-visual modalities, respectively. For audio-visual scenario, the decision-level fusion provided better performance in comparison to feature-level fusion. Humans convey their messages effectively by using both the audio and visual modalities. Each modality conveys unique information about different emotions. These modalities are highly correlated, and they complement each other in recognizing the human emotions (Hajarolasvadi & Demirel, 2020). For this reason, the audio-visual approach normally performs better as compared to unimodal approach.

Table 5: Audio-visual emotion recognition for decision-level fusion.

Attribute Evaluator	Search Method	Classification Accuracy (%)	
		Product Rule	Sum Rule
CFS	Best First	96.40	96.43
	Greedy Stepwise	68.47	96.60
Info Gain	Ranker	97.60	97.50
Gain Ratio	Ranker	96.80	97.51

Table 6: Confusion matrix of best classification accuracy obtained for decision-level fusion using the product rule.

Actual Emotion	Recognized Emotion						Accuracy (%) per Emotion
	A	E	F	H	N	S	
A = Anger	395	0	18	0	1	0	95.41
E = Excited	0	371	1	3	5	0	97.63
F = Frustration	2	0	748	0	0	0	99.73
H = Happy	0	7	0	220	6	3	93.22
N = Neutral	0	2	0	0	467	11	97.30
S = Sadness	0	1	0	0	8	509	98.70
Overall classification accuracy (%)							97.60

Table 7: Comparison between the audio, visual and bimodal emotion recognition results.

Modality			Classification Accuracy (%)			
			CFS		Info Gain	Gain Ratio
			Best First	Greedy Stepwise	Ranker	Ranker
`Audio			64.40	64.54	64.54	64.15
Visual			96.40	96.40	96.77	96.70
Audio-Visual	Feature Level		94.85	94.92	96.04	94.92
	Decision Level	Product	96.40	65.73	96.80	97.60
		Sum	96.43	96.60	97.51	97.50

Conclusion

In this research, automatic human emotion recognition was performed using the unimodal and bimodal approaches. The feature and decision levels fusion were used for the bimodal emotion recognition. The six emotion classes of the IEMOCAP database were used for experimentation. Features were selected using the CFS with best first and greedy stepwise search methods, info gain and gain ratio. The emotion recognition was performed using seven different classification techniques, i.e., Bayes Net, SVM, Meta Bagging, Rotation Forests, Functional Trees, Random Forests, and Random Trees.

The best performance of 64.54% was obtained for the audio modality using both the CFS (greedy stepwise search method) with rotation forest classifier, and info gain feature selection with meta bagging classifier. In the case of visual modality, the best result of 96.77% was obtained for the info gain feature selection with rotation forest classifier. The emotion recognition results improved for the bimodal scenario. The performance of decision-level fusion was better in comparison to feature-level fusion. The best classification result of 96.04% was obtained for feature-level fusion using the info gain feature selection with rotation forest classifier. The sum and product rules were used for decision-level fusion. For the product rule the best performance of 97.60% was obtained using the info gain feature selection, while for the sum rule the best result of 97.51% was obtained using the gain ratio feature selection.

Humans utilize both the audio and visual modalities to convey their messages effectively. Each modality delivers distinctive information about various emotions. These modalities seem to be correlated and they complement each other in recognizing human emotions. It is therefore the multimodal approach outperformed the unimodal approach. The future work includes utilizing the different audio-visual databases in various

languages. In addition, it will be interesting to use the face image data rather than the face marker data for the real-world applications.

References

Alluhaidan, A. S., Saidani, O., Jahangir, R., Nauman, M. A., & Neffati, O. S. (2023). Speech Emotion Recognition through Hybrid Features and Convolutional Neural Network. *Applied Sciences, 13*(8), 4750.

Araño, K. A., Gloor, P., Orsenigo, C., & Vercellis, C. (2021). When old meets new: emotion recognition from speech signals. *Cognitive Computation, 13*, 771-783.

Avots, E., Sapiński, T., Bachmann, M., & Kamińska, D. (2019). Audiovisual emotion recognition in wild. *Machine Vision and Applications, 30*(5), 975--985.

Azhagusundari, B., & Thanamani, A. S. (2013). Feature selection based on information gain. *International Journal of Innovative Technology and Exploring Engineering, 2*(2), 18-21.

Bhanusree, Y., Kumar, S. S., & Rao, A. K. (2023). Time-Distributed Attention-Layered Convolution Neural Network with Ensemble Learning using Random Forest Classifier for Speech Emotion Recognition. *Journal of Information and Communication Technology, 22*(1), 49-76.

Bota, P., Wang , C., Fred, A., & Silva, H. (2020). Emotion assessment using feature fusion and decision fusion classification based on physiological data: Are we there yet? *Sensors, 20*(17), 4723.

Breiman, L. (2001). Random forests. *Machine learning, 45*, 5-32.

Büchlmann, P., & Yu, B. (2002). Analyzing Bagging. *The Annals of Statistics, 30*(4), 927-961.

Busso, C., Bulut, M., Lee, C.-C., Kazemzadeh, A., Mower, E., Kim, S., et al. (2008). IEMOCAP: Interactive emotional dyadic motion capture database. *Language resources and evaluation, 42*, 335--359.

Chalapathi, M. V., Kumar, M. R., Sharma, N., & Shitharth, S. (2022). Ensemble Learning by High-Dimensional Acoustic Features for Emotion Recognition from Speech Audio Signal. *Security and Communication Networks, 2022*, 1-10.

Chang, C.-C., & Lin, C.-J. (2011). LIBSVM: a library for support vector machine. *ACM transactions on intelligent systems and technology (TIST), 2*(3), 1-27.

Chen, M., He, X., Yang, J., & Zhang, H. (2018). 3-D convolutional recurrent neural networks with attention model for speech

emotion recognition. *IEEE Signal Processing Letters, 25*(10), 1440--1444.

Ekman, P. (1993). Facial expression and emotion. *American Psychologist, 48*(4), 384--392.

Ekman, P., & Friesen, W. V. (1978). Facial action coding system: a technique for the measurement of facial movement. *Consulting Psychologists*.

Eyben, F., Wollmer, M., & Schuller, B. (2009). OpenEAR-Introducing the Munich Open-Source Emotion and Affect Recognition Toolkit. *International Conference on Affective Computing and Intelligent Interaction and Workshops* (pp. 1-6). IEEE.

Gama, J. (2004). Functional trees. *Machine learning, 55*, 219-250.

Hajarolasvadi, N., & Demirel, H. (2020). Deep emotion recognition based on audio-visual correlation. *IET Computer Vision, 14*(7), 517-527.

Hall, M. A. (1999). *Correlation-based feature selection for machine learning.* The University of Waikato.

Haq, S. U., Asif, M., Ali, A., Jan, T., Ahmad, N., & Khan, Y. (2015). Audio-visual emotion classification using filter and wrapper feature selection approaches. *Sindh University Research Journal-SURJ (Science Series), 47*(1), 67--72.

Haq, S., & Jackson, P. J. (2011). Multimodal Emotion Recognition. In *Machine Audition: Principles, Algorithms and Systems* (pp. 398-423). IGI Global.

Jain, A., Nandakumar, K., & Ross, A. (2005). Score normalization in multimodal biometric systems. *Pattern Recognition, 38*(12), 2270-2285.

Jain, Y. K., & Bhandare, S. K. (2011). Min max normalization based data perturbation method for privacy protection. *International Journal of Computer & Communication Technology, 2*(8), 45-50.

Khalil, R. A., Jones, E., Babar, M. I., Jan, T., Zafar, M. H., & Alhussain, T. (2019). Speech emotion recognition using deep learning techniques: A review. *IEEE Access, 7*, 117327--117345.

Khare, S. K., Blanes-Vidal, V., Nadimi, E. S., & Acharya, U. R. (2024). Emotion recognition and artificial intelligence: A systematic review (2014–2023) and research recommendations. *Information Fusion, 102*.

Liu, S., McGree, J., Ge, Z., & Xie, Y. (2016). *Computational and Statistical Methods for Analysing Big Data with Applications.* Elsevier.

Livingstone, S. R., & Russo, F. A. (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.

Ma, Y., Hao, Y., Chen, M., Chen, J., Lu, P., & Košir, A. (2019). Audio-visual emotion fusion (AVEF): A deep efficient weighted approach. *Information Fusion, 46*, 184--192.

Mannepalli, K., Sastry, P. N., & Suman, M. (2022). Emotion recognition in speech signals using optimization based multi-SVNN classifier. *Journal of King Saud University-Computer and Information Sciences, 34*(2), 384-397.

Mohan, M., Dhanalakshmi, P., & Kumar , R. S. (2023). Speech Emotion Classification Using Ensemble Models with MFCC. *Procedia Computer Science, 218*, 1857-1868.

Nie, W., Ren, M., Nie, J., & Zhao, S. (2020). C-GCN: Correlation based graph convolutional network for audio-video emotion recognition. *IEEE Transactions on Multimedia, 23*, 3793--3804.

Rodriguez , J. J., Kuncheva, L. I., & Alonso, C. J. (2006). Rotation Forest: A New Classifier Ensemble Method. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 28*(10), 1619-1630.

Samadiani, N., Huang, G., Cai, B., Luo, W., Chi, C.-H., Xiang, Y., et al. (2019). A review on automatic facial expression recognition systems assisted by multimodal sensor data. *Sensors, 19*(8), 1863.

Tan, L., Yu, K., Lin, L., Cheng, X., Srivastava, G., Lin, J. C.-W., et al. (2021). Speech emotion recognition enhanced traffic efficiency solution for autonomous vehicles in a 5G-enabled space--air--ground integrated intelligent transportation system. *IEEE Transactions on Intelligent Transportation Systems, 23*(3), 2830--2842.

The MathWorks Inc. (2024). MATLAB. Natick, Massachusetts, United States: The MathWorks Inc.

Thiam, P., Kestler, H. A., & Schwenker, F. (2020). Two-stream attention network for pain recognition from video sequences. *Sensors, 20*(3), 839.

Tzirakis, P., Trigeorgis, G., Nicolaou, M. A., Schuller, B. W., & Zafeiriou, S. (2017). End-to-end multimodal emotion recognition using deep neural networks. *IEEE Journal of selected topics in signal processing, 11*(8), 1301--1309.

Wang, Y., Guan, L., & Venetsanopoulos, A. N. (2012). Kernel cross-modal factor analysis for information fusionwith application to bimodal emotion recognition. *IEEE Transactions on Multimedia, 14*(3), 597--607.

Wang, Y., Wu, J., & Hoashi, K. (2019). Multi-attention fusion network for video-based emotion recognition. *International Conference on Multimodal Interaction.*

Witten, I. H., Frank, E., Hall , M. A., & Pal, C. J. (2016). *Data Mining: Practical Machine Learning Tools and Techniques.* Morgan Kaufmann.

Yuvaraj, R., Murugappan, M., Ibrahim, N. M., Sundaraj, K., Omar, M. I., Mohamad, K., et al. (2014). Detection of emotions in Parkinson's disease using higher order spectral features from brain's electrical activity. *Biomedical Signal Processing and Control, 14*, 108--116.

Zeng, Z., Tu, J., Pianfetti, B. M., & Huang, T. S. (2008). Audio--visual affective expression recognition through multistream fused HMM. *IEEE Transactions on multimedia, 10*(4), 570--577.

Zhang, S., Li, L., & Zhao, Z. (2012). Audio-visual emotion recognition based on facial expression and affective speech. *International Conference on Multimedia and Signal Processing* , (pp. 46-52). Shanghai.

Zhang, S., Zhang, S., Huang, T., Gao, W., & Tian, Q. (2017). Learning affective features with a hybrid deep model for audio--visual emotion recognition. *IEEE Transactions on Circuits and Systems for Video Technology, 28*(10), 3030--3043.

Zhao, G., Barnard, M., & Pietikainen, M. (2009). Lipreading with local spatiotemporal descriptors. *IEEE Transactions on Multimedia, 11*(7), 1254--1265.

Zhao, J., Mao, X., & Chen, L. (2019). Speech emotion recognition using deep 1D & 2D CNN LSTM networks. *Biomedical signal processing and control, 47*, 312--323.

Zhou, H., Du, J., Zhang, Y., Wang, Q., Liu, Q.-F., & Lee, C.-H. (2021). Information fusion in attention networks using adaptive and multi-level factorized bilinear pooling for audio-visual emotion recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 29*, 2617--2629.