Auditive emotion recognition for emphatic AI-assistants

Roswitha Duwenbeck[1*†] and Elsa Andrea Kirchner[1,2†]

[1]Systeme der Medizintechnik, Universität Duisburg-Essen, Bismarckstr. 81, Duisburg, 47057, NRW, Germany.
[2]Robotics Innovation Center, German Research Center for Artificial Intelligence (DFKI), Robert-Hooke-Strasse 1, Bremen, 28359, Bremen, Germany.

*Corresponding author(s). E-mail(s): roswitha.kressner@uni-due.de;
Contributing authors: elsa.kirchner@uni-due.de;
†These authors contributed equally to this work.

Abstract

This paper briefly introduces the Project "AudEeKA", whose aim is to use speech and other bio signals for emotion recognition to improve remote, but also direct, healthcare. This article takes a look at use cases, goals and challenges, of researching and implementing a possible solution. To gain additional insights, the main-goal of the project is divided into multiple sub-goals, namely speech emotion recognition, stress detection and classification and emotion detection from physiological signals. Also, similar projects are considered and project-specific requirements stemming from use-cases introduced. Possible pitfalls and difficulties are outlined, which are mostly associated with datasets. They also emerge out of the requirements, their accompanying restrictions and first analyses in the area of speech emotion recognition, which are shortly presented and discussed. At the same time, first approaches to solutions for every sub-goal, which include the use of continual learning, and finally a draft of the planned architecture for the envisioned system, is presented. This draft presents a possible solution for combining all sub-goals, while reaching the main goal of a multimodal emotion recognition system.

Keywords: emotion recognition, speech emotion recognition, multimodal emotion recognition, continual learning

1 Introduction

Rising patient numbers, paperwork, and a declining workforce are leading to tense working schedules in healthcare. In Germany, the average time a physician spends with each patient is reported to be about seven and a half minutes [7]. One of the many problems arising from these issues concerns the communication between patient and physician. Even though it is proven, that effective physician-patient communication leads to positive health outcomes [8], a stressful environment leads to reduced empathy [9] and ultimately reduces the quality of care which is provided, as empathy is a key-factor in effective patient encounters [10]. Furthermore, the rise of acceptance regarding AI-aided diagnoses [11], digitalization, and the promise of more accurate diagnoses and better therapeutic outcomes [12] mean that "soft factors" like empathy or stress management are increasingly neglected in medicine. The project "AudEeKA", short for "Auditive Emotionserkennung für empathische KI-Assistenten"

(eng..: auditive emotion recognition for emphatic AI-Assistants) aims to automate assessment of affective human states. One use case is supporting medical staff in estimating emotional and situational health of patients. This could help to indicate to the doctor when more time is needed with the patient, e.g. when the patient is in a state of emotional distress. Another use case is remote medical care, where assessing the emotional state of patients is more difficult, because a lacking or distorted overall impression of the patient can be a side effect of using technical possibilities for communication. The assessment of the emotional state becomes especially important, when psychological and medical care cannot be provided by telemedicine or even on-site, as it is the case in long time manned space flight. Because of the increased difficulty related to usage in space and on highly selected persons, it can be considered as a separate, but most important use-case, as the success of a mission can be coupled with emotional health.

The end-goal of "AudEeKA" is to reliable recognize emotions continuously in humans, by collecting multimodal, physiological signals. As the use cases require portable, edge computing devices or at least a solution which is able to run in environments which have little computing power and energy supply, this project will mainly look into solutions running on lightweight, low resource-devices. Furthermore, the computation should be fast enough to work in real time, to be used live in a conversation. By this the result could be an empathic AI-assistant, which classifies emotions and stress continuously in various environments.

To reach this goal, firstly speech signals should be used to classify emotions. After this first step, it is planned to combine emotion recognition with the stress-classification by using biosignals and maybe speech. Both, emotions and stress, should be traced, while possibly using and watching their synergy. Later on, more bio-signals should be included to gain accuracy in regards of classifying emotions. All of these sub-goals should naturally consider the resource consumption and computing time of the built algorithms and classificators.

2 Related Work

Emotion recognition (ER) is a topic of long-standing interest and has been researched in human-computer interaction [13], affective robotics [14], E-learning [15], automotive safety [16], customer-care [17] or healthcare [18].

Currently, most approaches for ER are based on speech-, text-, visual-, or physiological data [19]. Those who combine modalities, usually do so between speech and text(e.g. [20], [21], [22]), or speech and visual data (e.g. [23], [24], [25]). A good overview is given by the paper of Imani and Montazer, which, lists both unimodal and multimodal approaches [26]. There is not much existing work in exploiting the potential of combining speech with other biosignals for ER (like Blood-Volume-Pulse (BVP), Electromyography (EMG), Skin Conductance (SC), Respiration (RSP), Body temperature (Temp), Electroencephalography(EEG), Electrocardiogram (ECG), Heart Rate Variability (HRV), Electrodermal Activity (EDA)), even though literature advises otherwise [27]. Found approaches can be seen in table 1. Looking at table 1 it also gets apparent, that a lot of the found approaches use more computaionally expensive methods: While a Support Vector Machine (SVM), k-Nearest-Neighbours (k-NN) and Linear Discriminant Analysis (LDA) are less costly, a

Reference	Year	Modalities	Models
Kim, Andre [1]	2006	Speech, BVP, EMG, SC, RSP, Temp	LDA, k-NN, MLP
Chao, Linlin, et al. [2]	2015	Audio, Video, ECG, EDA	LSTM
Ranganathan, Chakraborty, Panchanathan [3]	2016	Face, Body, Voice and Physiologic Modalities (HR, ECG, RR interval, Breathing Rate, Posture, Activity level, Peak Acceleration)	SVM, DBN
Guo, Jiang, Shao [4]	2020	Speech, EEG, ECG	PNN, SVM, ELM
Bakhshi, Chalup [5]	2021	Audio signals, ECG, HRV	DNN
Wang, Wang, Yang, Zhang [6]	2022	EDA, SC, Speech, EEG	LDA, TCN, ELM, MLP

Table 1: Multimodal speech emotion recognition approaches

Multi Layer Perceptron (MLP), Long Short Time Memory (LSTM), Deep Belief Networf (DBN), Probabilistic Neural Network (PNN), Deep Neural Network (DNN), Temporal Convolutional Neural Network (TCN) or Extreme Learning Machine (ELM) are computational more expensive. Therefore, these last methods are not applicable to "AudEeKA". It should be noted that table 1 does not assert a claim for completeness. Table 1 also shows, that many approaches take visual signals, like features from face or posture, into account, which is often seen as a privacy-issue.
In order for all solutions for ER to work, annotated datasets are needed. A good overview of speech-datasets is given by Wani, Gunawan, Qadri, Kartiwi and Ambikairajah [28] or the Technical University of Munich [29]. Although the papers list speech-emotion-datasets in large numbers, there are multiple difficulties, when one wants to build models which can be used in real-life: There are not only many acted datasets, which do translate poorly into real-world, but also manifold different labels (basic emotions of varying numbers or the valence-arousal-scale). The same problem can be observed in datasets, which use physiological signals (as listed by Shu, Xie, Yang, Li, Li, Liao, Xu, Yang [30] or by Larradet, Niewiadomski, Barresi, Caldwell, Mattos [31]). Albeit the problem in physiological datasets lies not so much in acted emotions, as they are rarely acted, but in the sources which cause the emotions and do not fully resemble the real world. For example a lot of studies use the international affective picture system [32], while others are experimenting with the use of virtual reality games (e.g. [33]) or -systems (e.g. [34]) or music (e.g. [35]). Analogous to the speech-based datasets, multiple labelling-strategies are found in physiological datasets.

3 Goals and Challenges

Considering the manifold use-cases in Section 1, but also the related work in Section 2 it becomes clear that the end-goal mentioned in Section 1 can be divided in sub-goals to be ultimately combined, but with pitfalls and difficulties to be considered.

3.1 Speech emotion recognition

While there are already a variety of of thoroughly described solutions that practice speech-ER (SER), most of them have shortcomings with regard to the requirements of "AudEeKA". For instance there is a huge shift towards heavyweight deep-learning models, which cannot be considered in "AudEeKA", because of the low-resource-setting. Other problems arise, when using traditional machine-learning approaches: For example feature-sets have to be carefully chosen, but not in a way that is too specific just for the used dataset. Since the expression of affect is dependant on factors like culture [36], gender [37] or age [38] and the end-goal mentioned in Section 1 includes that the solution is working on a wide range of people (especially important in telemedicine settings) and in various environments, while also the resource-restriction has to be considered and would favour a minimal feature set.

To get a first understanding of the performance of different feature sets but also on inter-individual differences, first test results on the Berlin Database of Emotional Speech (Emo-DB), a speech database with a set of basic but acted emotions [39] are on hand in "AudEeKA". For classifications are done by a small MLP with hidden layer size of 340 and 32, logistic activation-function, maximum of 20 iterations and an initial learning rate of 0.0035 was applied. It was implemented with scikit-learn [40]. Parameters were choosen by experience. Used feature-sets were taken from Opensmile [41], namely emobase-functionals (988 features) and Compare2016-functionals (6373 features). For evaluation a Leave-One-Out-Crossvalidation (LOOCV) was used, in which the complete data of one subject was omitted as test-set in each iteration. With this LOOCV strategy, a more realistic test scenario was aimed for. In Fig.1 and 2 Boxplots with different statistical evaluations of the applied LOOCV can be seen. One conclusion here can be, that while working with MLPs, bigger feature sets are resulting in better recognition-rates and lesser outliers than smaller ones. That feature sets affect not only overall recognition rates but also emotion-specific recognition rates, can be seen in Fig.3 and 4. In Fig.3 and 4 are the emotion specific confusion matrices depicted, with the mean recognition accuracy percentage calculated

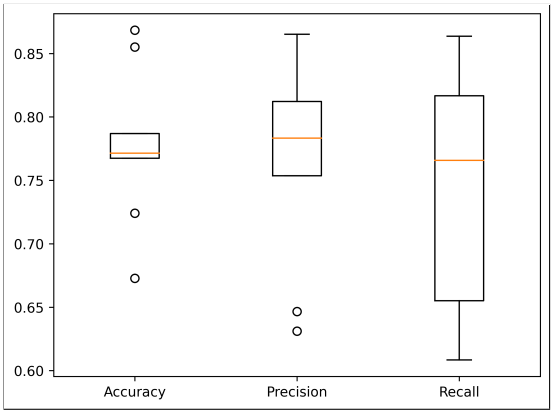

Fig. 1: Statistical values of MLP-classifier with emobase feature set on Emo-DB with LOOCV

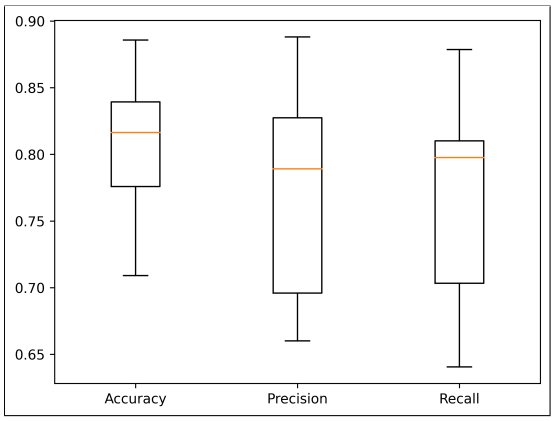

Fig. 2: Statistical values of MLP-classifier with Compare2016 feature set on Emo-DB with LOOCV

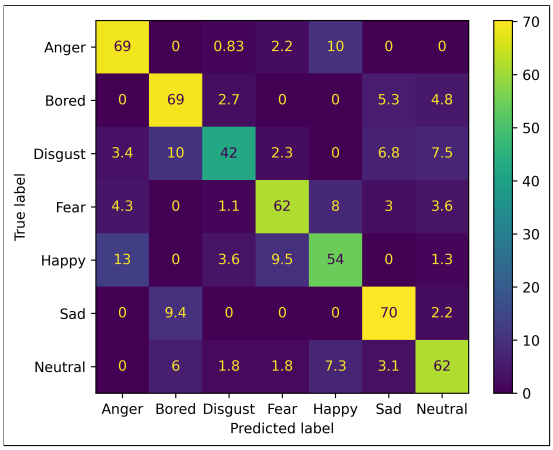

Fig. 3: Averaged Confusion Matrix of MLP with emobase feature set on Emo-DB with LOOCV

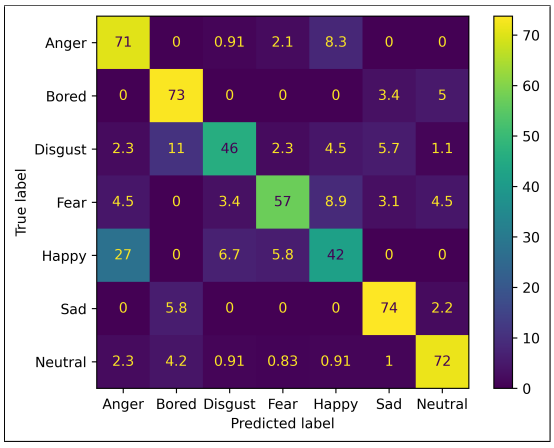

Fig. 4: Averaged Confusion Matrix of MLP with Compare2016 feature set on Emo-DB with LOOCV

for every emotion. It is apparent that, in general, there are emotions that get mixed up more often than others. When looking at Fig 1 and 2 it also becomes clear that the recognition value varies strongly from person to person. Ultimately it is unrealistic, that there is a "one-size-fits-all"-model, which works well on an individual in the wild. Not only the inter-individual differences are remarkable, there are also few datasets which record speech in the wild (also seen in Section 2).

Furthermore, the professional environment, as in manned space-flight or healthcare leaves little room for showing emotions. Oftentimes there are pre-existing speech-patterns, which further complicate the SER. To overcome these barriers, another sub-goal of the project is to create a new dataset. An additional sub-goal is to implement an approach that allows continuous learning, as it could help to eliminate inter-individual problems of (S)ER.

3.2 Stress detection and classification

Stress and emotions have undoubtedly a close relationship. Most emotion-models do not consider stress an emotion. This includes most theories

which use the term “basic emotions” [42], or Russel’s circumplex of emotions [43], although the fact if a person is stressed, certainly is an essential context for evaluating the importance or severity of an expressed affect. Especially long-term stress, which “can potentially cause those cognitive, emotional and behavioral dysfunctions” [44]. To add context, it could be important to monitor stress. In the end, the detection of stress could not only be beneficial to the accuracy of SER, but also for other health-applications and a general monitoring of stress.

3.3 Emotion detection from physiological signals

One lesson from subsection 3.1 is that ER faces numerous challenges. To gain further insights and increase accuracy of the predictions using more input-modalities could be beneficial. At the moment a minimalistic use of modalities is anticipated for “AudEeKA”. Not depicted is a possible system or method for user input for continual learning and to retrain the model(s), since the exact strategy for continual learning is not fully elaborated. Because continual learning is considered crucial, it has to be thoroughly researched in literature and by own implementation of prototypes. The research and implementation of prototypes for continual learning can be considered part of the continual learning related sub-goal.

3.4 Combining all approaches

In order to unite all planned classifications, namely SER, stress-classification and biophysiological-signal based ER, into one, fusion approaches must be applied. As there are currently to the authors knowledge no datasets, which fit the needs of this particular approach including not only stress, but also speech and other biosignals to ER, it is planned to train all models alone, on suitable datasets. In future work we plan to measure the performance of the fused models against the specifically created dataset. The planned architecture of the approach can be seen in Fig. 5. Different results from different classifications are planned to be taken into account for the other classifications, as there is maybe a synergy. To save space, the ER in Fig. 5 is depicted as one model and output. Generally, the design in Fig. 5 sets its main focus on feature-fusion. Due to the multiple types of sensors to be used, data-level-fusion would hardly be possible, as is usually aims to combine multiple homogenous data sources [45]. Furthermore, feature-level fusion has proven to be successfull in ER with other modalities (e.g. [46], [47], [48]). Since this is also the case for decision-level fusion (e.g. [48], [49] or [50]), the proposed architecture can easily be converted to fit an decision-level fusion approach. Also, the sub-goals defined in Section 1 are reflected in Fig. 5. It can be seen, that the tasks of emotion- and stress-classification are executed and combined, to obtain a better classification result.

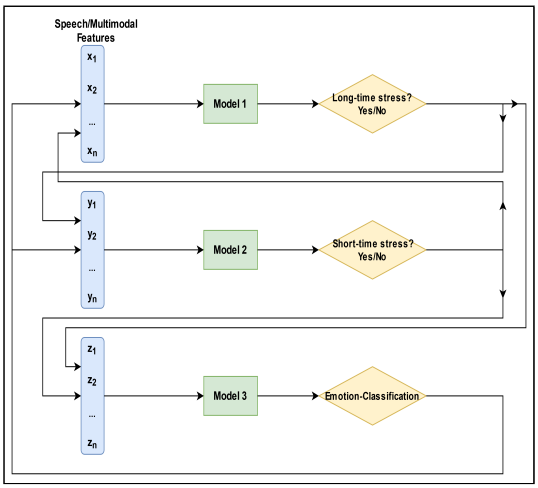

Fig. 5: Planned architecture of the emotion-recognition system

4 Conclusion

The Project “AudEeKA” is addressing a wide scope of challenges, but if executed correctly also major benefits for healthcare in general. This article describes only the first steps and provides summary insight into the initial considerations and the resulting future developments and research questions. To prove the suitability of the concepts presented here, a variety of implementations and tests must be performed on different datasets, including the creation of a dedicated dataset. Required tests mainly aim at the suitability for online usage and usage in the real world or in scenarios which come close to the use cases.

This results in test-scenarios which involve various noisy backgrounds, people of different ages, gender and cultural backgrounds. The tests also have to consider fast but nevertheless accurate classification results and aim to obtain these classification results with minimal use of resources. To reach the best possible results, different classifiers, feature sets and combination possibilities (feature- or decision-fusion) have to be implemented and compared.

Acknowledgments. This work is funded by the Federal Ministry for Economic Affairs and Climate Action and the German Aerospace Center [grant number 50RP2260A]. We acknowledge support by the Open Access Publication Fund of the University of Duisburg-Essen.

Declarations

The authors declare that they have no conflict of interest.

References

[1] J. Kim and E. André, "Emotion recognition using physiological and speech signal in short-term observation," in *Perception and Interactive Technologies: International Tutorial and Research Workshop, PIT 2006 Kloster Irsee, Germany, June 19-21, 2006. Proceedings*, pp. 53–64, Springer, 2006.

[2] L. Chao, J. Tao, M. Yang, Y. Li, and Z. Wen, "Long short term memory recurrent neural network based multimodal dimensional emotion recognition," in *Proceedings of the 5th international workshop on audio/visual emotion challenge*, pp. 65–72, 2015.

[3] H. Ranganathan, S. Chakraborty, and S. Panchanathan, "Multimodal emotion recognition using deep learning architectures," in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1–9, 2016.

[4] H. Guo, N. Jiang, and D. Shao, "Research on multi-modal emotion recognition based on speech, eeg and ecg signals," in *Robotics and Rehabilitation Intelligence: First International Conference, ICRRI 2020, Fushun, China, September 9–11, 2020, Proceedings, Part I 1*, pp. 272–288, Springer, 2020.

[5] A. Bakhshi and S. Chalup, "Multimodal emotion recognition based on speech and physiological signals using deep neural networks," in *Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, January 10–15, 2021, Proceedings, Part VI*, pp. 289–300, Springer, 2021.

[6] Q. Wang, M. Wang, Y. Yang, and X. Zhang, "Multi-modal emotion recognition using eeg and speech signals," *Computers in Biology and Medicine*, vol. 149, p. 105907, 2022.

[7] C. Winnat, "Deutsche aerzte nehmen sich rund sieben minuten zeit pro patient," Nov 2017.

[8] "Effective physician-patient communication and health outcomes: a review.," *CMAJ: Canadian medical association journal*, vol. 152, no. 9, p. 1423, 1995.

[9] J. P. Nitschke and J. A. Bartz, "The association between acute stress & empathy: A systematic literature review," *Neuroscience & Biobehavioral Reviews*, p. 105003, 2022.

[10] D. C. Dugdale, R. Epstein, and S. Z. Pantilat, "Time and the patient-physician relationship," *Journal of general internal medicine*, vol. 14, p. S34, 1999.

[11] K. Budde, T. Dasch, E. Kirchner, U. Ohliger, M. Schapranow, T. Schmidt, A. Schwerk, J. Thoms, T. Zahn, and K. Hiltawsky, "Künstliche intelligenz: Patienten im fokus," *Dtsch Arztebl*, vol. 117, no. 49, pp. A–2407, 2020.

[12] L. S.-D. P. L. Systeme, "Lernende systeme im gesundheitswesen: Grundlagen, anwendungsszenarien und gestaltungsoptionen," *Bericht der AG Gesundheit, Medizintechnik, Pflege*, 2019.

[13] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. G. Taylor, "Emotion recognition in human-computer interaction," *IEEE Signal*

processing magazine, vol. 18, no. 1, pp. 32–80, 2001.

[14] A. Austermann, N. Esau, L. Kleinjohann, and B. Kleinjohann, "Prosody based emotion recognition for mexi," in *2005 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 1138–1144, IEEE, 2005.

[15] H. Altun, "Integrating learner's affective state in intelligent tutoring systems to enhance e-learning applications," *GETS 2005*, vol. 3, p. 1, 2005.

[16] C. L. Lisetti and F. Nasoz, "Using noninvasive wearable computers to recognize human emotions from physiological signals," *EURASIP Journal on Advances in Signal Processing*, vol. 2004, pp. 1–16, 2004.

[17] L. Devillers, L. Lamel, and I. Vasilescu, "Emotion detection in task-oriented spoken dialogues," in *2003 International Conference on Multimedia and Expo. ICME'03. Proceedings (Cat. No. 03TH8698)*, vol. 3, pp. III–549, IEEE, 2003.

[18] D. Tacconi, O. Mayora, P. Lukowicz, B. Arnrich, C. Setz, G. Troster, and C. Haring, "Activity and emotion recognition to support early diagnosis of psychiatric diseases," in *2008 Second International Conference on Pervasive Computing Technologies for Healthcare*, pp. 100–102, IEEE, 2008.

[19] A. Saxena, A. Khanna, and D. Gupta, "Emotion recognition and detection methods: A comprehensive survey," *Journal of Artificial Intelligence and Systems*, vol. 2, no. 1, pp. 53–79, 2020.

[20] M. R. Makiuchi, K. Uto, and K. Shinoda, "Multimodal emotion recognition with high-level speech and text features," in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 350–357, 2021.

[21] L. Pepino, P. Riera, L. Ferrer, and A. Gravano, "Fusion approaches for emotion recognition from speech using acoustic and text-based features," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6484–6488, 2020.

[22] N.-H. Ho, H.-J. Yang, S.-H. Kim, and G. Lee, "Multimodal approach of speech emotion recognition using multi-level multi-head fusion attention-based recurrent neural network," *IEEE Access*, vol. 8, pp. 61672–61686, 2020.

[23] L. Schoneveld, A. Othmani, and H. Abdelkawy, "Leveraging recent advances in deep learning for audio-visual emotion recognition," *Pattern Recognition Letters*, vol. 146, pp. 1–7, 2021.

[24] L.-A. Perez-Gaspar, S.-O. Caballero-Morales, and F. Trujillo-Romero, "Multimodal emotion recognition with evolutionary computation for human-robot interaction," *Expert Systems with Applications*, vol. 66, pp. 42–61, 2016.

[25] A. I. Middya, B. Nag, and S. Roy, "Deep learning based multimodal emotion recognition using model-level fusion of audio–visual modalities," *Knowledge-Based Systems*, vol. 244, p. 108580, 2022.

[26] M. Imani and G. A. Montazer, "A survey of emotion recognition methods with emphasis on e-learning environments," *Journal of Network and Computer Applications*, vol. 147, p. 102423, 2019.

[27] S. G. Koolagudi and K. S. Rao, "Emotion recognition from speech: a review," *International journal of speech technology*, vol. 15, pp. 99–117, 2012.

[28] T. M. Wani, T. S. Gunawan, S. A. A. Qadri, M. Kartiwi, and E. Ambikairajah, "A comprehensive review of speech emotion recognition systems," *IEEE Access*, vol. 9, pp. 47795–47814, 2021.

[29] T. U. Muenchen, "Eight emotional speech databases used - tum."

[30] L. Shu, J. Xie, M. Yang, Z. Li, Z. Li, D. Liao, X. Xu, and X. Yang, "A review of emotion recognition using physiological signals," *Sensors*, vol. 18, no. 7, p. 2074, 2018.

[31] F. Larradet, R. Niewiadomski, G. Barresi, D. G. Caldwell, and L. S. Mattos, "Toward emotion recognition from physiological signals in the wild: approaching the methodological issues in real-life data collection," *Frontiers in psychology*, vol. 11, p. 1111, 2020.

[32] P. J. Lang, M. M. Bradley, B. N. Cuthbert, *et al.*, "International affective picture system (iaps): Technical manual and affective ratings," *NIMH Center for the Study of Emotion and Attention*, vol. 1, no. 39-58, p. 3, 1997.

[33] P. Merkx, K. P. Truong, and M. A. Neerincx, "Inducing and measuring emotion through a multiplayer first-person shooter computer game," in *Proceedings of the Computer Games Workshop*, 2007.

[34] W. Zhang, L. Shu, X. Xu, and D. Liao, "Affective virtual reality system (avrs): design and ratings of affective vr scenes," in *2017 International Conference on Virtual Reality and Visualization (ICVRV)*, pp. 311–314, IEEE, 2017.

[35] J. Kim and E. André, "Fusion of multichannel biosignals towards automatic emotion recognition," *Multisensor Fusion and Integration for Intelligent Systems*, vol. 35, no. Part 1, pp. 55–68, 2009.

[36] D. Matsumoto, "Ethnic differences in affect intensity, emotion judgments, display rule attitudes, and self-reported emotional expression in an american sample," *Motivation and emotion*, vol. 17, no. 2, pp. 107–123, 1993.

[37] L. R. Brody, "On understanding gender differences in the expression of emotion," *Human feelings: Explorations in affect development and meaning*, pp. 87–121, 1993.

[38] R. W. Levenson, L. L. Carstensen, W. V. Friesen, and P. Ekman, "Emotion, physiology, and expression in old age.," *Psychology and aging*, vol. 6, no. 1, p. 28, 1991.

[39] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, B. Weiss, *et al.*, "A database of german emotional speech.," in *Interspeech*, vol. 5, pp. 1517–1520, 2005.

[40] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[41] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM international conference on Multimedia*, pp. 1459–1462, 2010.

[42] J. L. Tracy and D. Randles, "Four models of basic emotions: A review of ekman and cordaro, izard, levenson, and panksepp and watt," *Emotion review*, vol. 3, no. 4, pp. 397–405, 2011.

[43] J. A. Russell, "A circumplex model of affect.," *Journal of personality and social psychology*, vol. 39, no. 6, p. 1161, 1980.

[44] A. Mariotti, "The effects of chronic stress on health: new insights into the molecular mechanisms of brain–body communication," *Future science OA*, vol. 1, no. 3, 2015.

[45] T. Gao, J.-Y. Song, J.-Y. Zou, J.-H. Ding, D.-Q. Wang, and R.-C. Jin, "An overview of performance trade-off mechanisms in routing protocol for green wireless sensor networks," *Wireless Networks*, vol. 22, pp. 135–157, 2016.

[46] H. Gunes and M. Piccardi, "Affect recognition from face and body: early fusion vs. late fusion," in *2005 IEEE International Conference on Systems, Man and Cybernetics*, vol. 4, pp. 3437–3443 Vol. 4, 2005.

[47] D. Hazarika, S. Gorantla, S. Poria, and R. Zimmermann, "Self-attentive feature-level fusion for multimodal emotion detection," in *2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, pp. 196–201, 2018.

[48] W.-L. Zheng, B.-N. Dong, and B.-L. Lu, "Multimodal emotion recognition using eeg and eye tracking data," in *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 5040–5043, 2014.

[49] S. Sahoo and A. Routray, "Emotion recognition from audio-visual data using rule based decision level fusion," in *2016 IEEE Students? Technology Symposium (TechSym)*, pp. 7–12, 2016.

[50] K.-S. Song, Y.-H. Nho, J.-H. Seo, and D.-s. Kwon, "Decision-level fusion method for emotion recognition using multimodal emotion recognition information," in *2018 15th International Conference on Ubiquitous Robots (UR)*, pp. 472–476, 2018.