Expose & Work Plan

Master of Science in Informatics

The Role of Physiological Signals in Multimodal Emotion Recognition Solutions in Autonomous Driving

by

Simge Özcan

Supervisor: Sina Shafaei
Handed in: February 5, 2021

1. Introduction and Background

As a result of recent advancements in the field of Computer Science, machines and humans increasingly engage in more dynamic ways and this interaction is yet to increase more. In order to improve the quality of interaction, it is vital that machines have an understanding of humans' emotional states. Affective Computing is a field of study which emerged as a result of this need. It is a multidisciplinary field which focuses on better understanding how humans recognise, interpret and simulate emotional states.

Affective Computing has applications in many areas such as healthcare, driving assistance and education. We focus on particularly autonomous driving context to improve the safety and performance through an automated driving assistance system (ADAS). The system could be capable of warning the user if he is sleepy, unconscious or unhealthy to drive, lowering the speed or stopping the car if necessary, towards a more safe and secure driving experience.

There are several techniques used to recognize the emotional state of an individual. The most common techniques used in the autonomous driving context are analysis of facial expressions, voice recognition, driver behaviour, eye tracking and analysis of physiological signals. The goal of our research is to investigate the importance of the physiological signals in recognition process and to develop a multimodal emotion recognition system for the cabin of a vehicle that is focused on physiological signals.

Having a multimodal recognizer in the cabin environment is vital to have continuous accurate detection of the current emotion of the driver. For instance, a modality that depends purely on camera signal could easily be deceived when the camera vision is obstructed or simply when the camera is malfunctioning. Thus, depending on a sensor data from one device is not reliable even though the accuracy could be high. Use of several modalities and sensor fusion is highly important.

Physiological signals plays a crucial role when it comes to emotions. Emotions are not only expressed by visible inputs such as facial expressions or gestures but also invisible inputs such as heart rate, breathing rate, body temperature, and sweating level. Acquiring these signals in a cabin environment is not easy and applicable due to the complexity of the hardware. Also, most sensors used in this technique are intrusive way of detecting the emotions. Therefore, the presence of such tools could also affect the current emotions of the driver. Thus, we should use those signals which could be acquired in the least intrusive way and is applicable to a cabin environment.

2. Related Work

Since we will investigate the role of physiological signals in multimodal emotion recognition, it is important to understand this field in a broader perspective. In this section we present an overview the modalities and techniques used to recognize emotions with a focus on applicability in the autonomous driving domain.

2.1. Modalities for Emotion Recognition

2.1.1. Audio-based

Audio-based approaches are one of the most researched and applicable models. Since emotions like anger, irritation, and nervousness are strongly correlated to simple speech features like volume, energy or pitch, audio-based approaches found to be especially good for the recognition of these driver states [6]. However, driver states such as sleepiness could not be reliably recognized with the speech signal as a sleepy driver is not tend to talk much. One of the great advantages of the speech modality is low hardware costs and giving less apparent observation feeling to the driver [4]. This makes the audio-based approaches more applicable from the user acceptance point of view. On the other hand, the user can control how much emotion is shown, which of course is a disadvantage for a constant and reliable driver monitoring. Audio information is not continuously present if the driver does not constantly speak. For instance, the previously mentioned sleepiness state is one of the greatest examples of this situation. Furthermore, studies have shown that open microphone settings and language differences make recognition tasks depending on audio challenging [19]. As all audio captured by the microphone may not be relevant for the recognition task and emotions have found to be highly related to the language spoken. The mean accuracy of the studies varies depending on the language. Therefore, language dependant studies result in better estimations [14].

2.2. Behavioural-based

The final approach is the use of driving style as a modality for emotion recognition. It exploits features like steering wheel, acceleration, lane-keeping and reacting. Since the driving style that is relevant to these features and the active state of the driver are highly correlated, it could be utilized for in-vehicle emotion recognition. For level 3 and level 4 autonomous driving, it could seem a bit useless however, the steering wheel can still exist and the driver is likely to drive the car. It still worth mentioning as the results are quite promising. One of the advantages of this approach is that it is quite obvious from

the emotional point of view. Further, it is less costly as it does not require any extensive hardware compared to physiological approaches. However, it is not investigated very intensely so far [4].

2.2.1. Camera-based

As the face is considered to be the major input resource for recognizing the emotions, video and image based approaches are the most focused one among the other approaches. This approach exploits the features like eye tracking, head movement, and facial expressions. Recognizers which use visual information have found to be good at detecting 6 basic emotions like anger, sadness, happiness, disgust, fear, irritation, and surprise. It is seen to be especially good for detecting the interest level of the driver compared to the audio modalities [4]. In contrast to speech, one advantage of the visual approaches is that the visual data is continuously present. There- fore, emotions could not be hidden or controlled by the driver. However, this also gives an increased observation feeling to the driver which is a disadvantage from a user acceptance point of view.

2.3. Physiology-based

Besides from a pattern recognition point of view, more extensive approaches are to collect and analyze the physiological signals, which include the electroencephalogram (EEG), body temperature (BT), electrocardiogram (ECG), electromyogram (EMG), electrodermal activity (EDA), respiration rate (RSP), etc. One advantage of this approach is that it provides reliable and better accuracy as the data is directly coming from the user's body and is highly correlated to emotions such as excitement, anger, and nervousness. Depending on the type of signal and the type of hardware that will be used for the recognition task, the hardware costs of physiological measurements could be affordable. However, it would be a disadvantage to wear devices that give an apparent feeling of being watched from a user acceptance point view. [4]. This consequently also affects the user's current emotion and behaviours. Further, working with these kinds of signals is highly challenging. For instance, EEG data can be noisy and diverse. Adequate amount of research has been done in this field which exploits several different types of physiological data. In the following section, we present some of the previously done work utilizing different physiological signals.

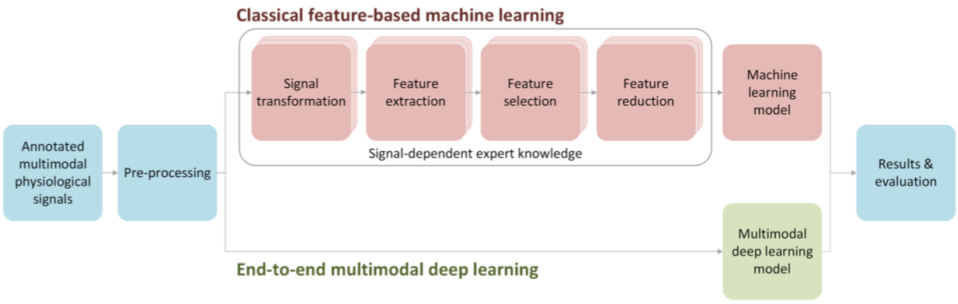

Figure 1: Comparison of feature-based ML methods and end-to-end DL methods which is applicable to emotion recognition from physiological signals problem [8]

2.3.1. Machine Learning

So far, most of the studies focusing on physiological signals have applied various feature-based machine learning (ML) approaches which includes extracting the valuable features from signals and processing those features based on some expert knowledge in order to train classifiers for recognizing emotions. Figure 1

Khezri et al. [11] combined EEG with GSR to recognize six basic emotions via k-nearest neighbors (kNN) classifiers. Yin et al. [25] developed an approach for emotion recognition from physiological signals using an ensemble of autoencoders. They used features from seven physiological sensors: EEG, EDA, EOG, EMG, ECG, BVP, and the respiration rate sensor as an input. In general, the methods based on EEG data usually outperform the ones based on other data [20]. It probably comes from the fact that EEG provides a more direct channel to one's mind.

In the field of stress detection using physiological sensors and ML are Healey and Picard [7], who proposed a quite accurate stress detection system in 2005. It achieved an accuracy as high as 97% when tested in a constrained real-life scenario, i.e., subjects driving a car. They used features extracted from ECG, EMG, EDA, and respiration to feed the input of the system. Even though the presented system was obtrusive, it confirmed that stress detection was possible in a real-life scenario even in 2005.

All of these systems for emotion and stress recognition are based on features extracted from signal data provided by the wearable sensors. The advanced domain-dependent signal processing and transformation (e.g., wavelet/Fourier transform, heart rate extraction from BVP, and spectral analysis), various feature extraction (including signal morphology and statistical and nonlinear measures), as well as feature selection and reduction are the most important and most challenging steps in the overall processing pipeline. Figure 1

2.3.2. Deep Learning

Deep Learning is a class of ML algorithms that use layers of nonlinear processing units, which are typically neurons. The first layer receives the input data, and each successive layer accepts the output from the previous layer as its input. A common advantage of the DL models based on Convolutional Neural Networks (CNNs) and Long Short Term Memory (LSTMs) is that they can learn directly from the raw data, thus avoiding the need for signal and feature processing. Figure 1

In a study of the valance level during walking in the city center, participants filled Self-Assessment Mankin (SAM) questionnaires. [10] The collected data, including heart rate, EDA, body temperature, and motion data, was put into three different end-to-end DL architectures. Multi-Layer Perceptron achieved an F1-score of 0.63, the CNNs achieved 0.71, and the CNN-LSTM achieved 0.874. Another approach to end-to-end DL was presented by Li at al. [12]. Firstly, they transformed each raw signal into a spectrogram. Later, these spectrograms were fed into an attention-based bidirectional LSTM network and then to an unspecified DNN. Their network achieved an F1-score of 0.72 for binary arousal recognition and 0.70 for binary valence evaluation. Qiu at al. [16] proposed Correleted Attention Networks for emotion recognition using bidirectional Gated Recurrent Units (GRUs), a Canonical Correlation Layer, a Signal Fusion Layer, an Attention Layer, and a Classification Layer. This architecture was tested on three datasets: SEED, SEED IV, and DEAP. Their framework achieved higher accuracy than feature-based SVM, although details about the features utilized were not provided. Additionally, CNNs were used on the MAHNOB-HCI dataset to achieve better accuracies than those found using the methods based on feature extraction [18].

2.4. Multimodal Fusion

Although approaches exploiting single modalities seem to achieve above baseline accuracy, a single modality could not be fully reliable in the real world environment. In an in-vehicle environment, it is vital to have redundancy by utilizing data from different sensors. This requires processing data from more than one modality and find relations between modalities, just like intelligent beings do in an naturally noisy and multimodal world. Many ML-based fusion approaches have been proposed to handle multimodal information for classification tasks. The most widely used methods are data level fusion (or early integration) and decision level fusion (or late integration). Data level fusion integrates low-level features from each modality by correlation, which potentially accomplishes better task, but has difficulty in temporal synchronization among various input sources. Decision level fusion obtains unimodal decision values and integrates

them to obtain the final decision. Although the late fusion ignores some low-level interactions between modality, it allows easy training with more flexibility, and simplicity to make predictions when one or more of modalities are missing. The hybrid (or mid-level) fusion attempts to exploit the advantages of both early and late fusion in a common framework [15].

[9] proposed an emotion recognition framework for semi-autonomous vehicles using a k-nearest neighbors algorithm (k-NN) as the classifier.. Although it is one of the simplest machine learning classifiers, the results observed to be much better than most of the other state-of- the-art classifiers. They focused on an approach to determine the driver's state of mind from these 9 different emotions. For the evaluation part, Database for Emotion Analysis using Physiological Signals (DEAP) and the MAHNOB-HCI Tagging Database which include records for EEG signals for each stimulus were selected and they were time-synchronized with other physiological variables as well as videos of the facial expressions. Tzirakis et. al proposed a multimodal system that operates on the raw signal. To perform an end-to-end spontaneous emotion prediction task from speech and visual data LSTMs were used. They fused visual and speech modalities and to speed up the training, unimodal networks were pretrained separately. The multimodal model greatly outperformed the other models. [22] Their work is open to be improved by adding other modalities to the same fusion architecture. Furthermore, according to [5] DL approaches require more modalities in order to perform better. They compared ML and DL methods exploiting physiological and visual signals, namely, nEDA, pEDA, HR, BR and eye tracking data.

3. Research Questions and Challenges

In this work, we will explore multimodal emotion recognition focusing on physiological signals with a purpose of applicability in an in-vehicle environment. Based on the literature review and related work covered, we set our research direction to attempt to answer the following primary research questions:

- Which physiological signals could be used for in-cabin environment in order to recognize emotions in a non-invasive way?
- How to map the features of the physiological signals to fit the different emotion states?
- How pre-processing phase on data could enhance the DL performance when applied before model training?
- What common characteristics do DL architectures that perform well on physiological data have?
- How is the accuracy of DL approaches for emotion recognition from physiological signals, especially when combined with the signal pre-processing techniques before model training?
- How does fusing camera-based approaches with physiological signals enhance the performance and what are the robust fusion solutions?

One of the main challenges that we could face during this research is working with raw physiological data. Raw data could have a high degree of noise due to the events occurring during the data acquisition process such as subject movement unstable environment temperature and humidity, and other non-related user movements that could occur in-vehicle environment. Moreover, there might be missing data caused by the disconnection of the sensor from the subject's body. Dealing with missing data and finding appropriate pre-processing techniques for each different physiological signal that will be utilized in order to reduce the noise to the minimum and make the most out of it will be the biggest challenges.Another possible challenge that we could face is to find a suitable fusion approach that would correlate the different modalities accurately. Multimodal emotion recognition necessitates the fusion of the modal features extracted from the raw signals. It is observed that performance of the fusion technique depends on the number of modalities, features extracted, types of classifiers, and the dataset used in the analysis. Thus, It is still unclear which fusion techniques outperform the others. [13].

4. Proposed Approach and Methodology

In this chapter, we outline the proposed approach and the methodology through which we attempt to address our main research questions. In order to investigate the role of physiological signals in a multımodal in-vehicle emotion recognition, we propose developing an end-to-end DL architecture for the classification task. We will feed this network with the pre-processed raw sensor data to discover useful patterns. Later, we will integrate this modality in a multimodal fusion model. At the end we would find out the appropriate pre-processing methods for physiological signals, the accuracy and flexibility level of multimodal systems for in-cabin environment, and compare the benchmarks and limitations of uni/multi-modal solutions focusing on physiological signals.

Overall main steps required in the development of a ML system for emotion recognition includes:

- Data Collection
- Signal Pre-processing
- Feature Engineering / Dimensionality Reduction
- Classification
- Validation and Evaluation

We will follow these steps and finally also fuse the developed modality into multimodal classification architecture. Next, we will elaborate which methodologies we will use in these steps according to our proposed approach.

4.1. Data Collection and Dataset Selection

First of all, we will make use of an existing data sources from Taamneh et al. [21] which consist of multiple modalities of affective information. It contains multimodal signals acquired in a controlled experiment on a driving simulator that involves n = 68 volunteers that drove the same highway under four different conditions: no distraction, cognitive distraction, emotional distraction, and sensorimotor distraction. The signals included in the database are speed, acceleration, brake force, steering, and lane position signals, perinasal perspiration (PER-EDA), palm EDA (P-EDA), heart rate (HR), breathing rate (BR), facial expression signals, and eye tracking data. Primary reasons of this dataset choice are that the physiological signals were obtained on a driving simulator with unobstructive sensors and the dataset also includes data for other modalities such as camera

based or behavioral based. These factors perfectly meet our study goals. Thus, the data collection process is irrelevant for us.

4.2. Pre-processing

Since, we are focusing on physiological signals, we will be dealing with a raw sensor signals which are noisy and exposed to external interference such as subject movement, disconnection, unstable environment temperature and missing data. Signal pre-processing step is vital for our research and it will be applied to the raw signals as a first step. Generally pre-processing step consists of: synchronisation of the different sensor's signals; removal of data-loss and null values; filtering, noise, and outlier removal. Most common pre-processing techniques used on ECG, EDA, HR and BR signals are butterworth filter, low-pass filter, smoothing, downsampling, winsorization, and normalization [1]. Selection of technique and and related parameters depends on the data and sensors used to collect the data. Therefore, at this step we will use trial and error method to find out the right techiques that fits to the data. As we will be using feature independent ML methods, we need little to no requirement of handcrafted features. Thus, feature extraction step is also irrelevant. We will only use pre-processing techniques that are not related to the specific domain knowledge and to a corresponding ML technique. This step is also vital for our research as one of our research questions is directly related to analysing the impacts of the pre-processing on DL performance.

4.3. Classification

Next step is to develop an end-to-end DL architecture that can learn from several sensors simultaneously, mostly inspired from Gjoreski et al. [3] They compared ten end-to-end DL architectures that are specialized for time-series classification in their study. Their work provides a solid base for the further improvements of the DL architectures and analyzing the influence of the pre-processing techniques. We will utilize these architectures by considering our chosen dataset and signals. Three best performing DL architectures according to their experimentation on four different datasets are Spectrotemporal Residual Network (Stressnet) [5], Residual Network (Resnet) [24], and Fully Convolutional Network (FCN) [23]. These architectures were take from corresponding cited works and enhanced to learn from multilevel physiological signals. Most accurate results were obtained from WESAD dataset which focuses on stress detection and with ECG, EDA, BCP and RSP. As we will also plan to use the same signals with a similar classification purposes, utilizing these architectures for our dataset might give promising results. Additionally, Stresnet were originally proposed for driver distraction detection [5].

4.4. Validation and Evaluation

In the validation and evaluation step, we will compare the results of architecture on different modes of pre-processed data to analyze the influence of different signal pre-processing techniques or the impact of signal selection. By conducting several experiments, one can also compare and identify the characteristics of DL architectures that can learn better from physiological signals.

4.5. Fusion

Finally, we will develop a sensor fusion approach to combine the chosen modality or modalities of behavior with our physiological signals-based modality to enhance the robustness or accuracy of the model. For this, two different approach could be followed. First is to make use of a previously developed multimodal fusion approach in the Chair of Robotics, Artificial Intelligence and Embedded Systems which achieved scores up to %54 by fusing behavioral, physiological and visual modalities. This design was modular and extensible which could support possible new networks. Therefore, we can easily analyze the effects on the accuracy by integrating our best performing physiological signals modality to this multimodal approach. Another approach is to follow a robust deep learning architecture for multimodal classification [2], EmbraceNet, to developed physiology-based modality with other modalities, mostly by reusing the modalities developed previously. EmbraceNet is a novel approach for multimodal integration for DL models, which provides good compatibility with any network structure and smooth handling of missing data. This integration approach could give satisfying results with our chosen dataset as it has considerable amount of missing data.

A. Outline Draft

1. Chapter one
    - Introduction
2. Chapter two
    - Background and related work
3. Chapter three
    - Methodology
4. Chapter four
    - Evaluation and Results

B. Time and Tasks Planning

Detailed planning of the individual steps, tasks in months.

Monat	Aufgabe	Meilenstein
January-February	Literature research Expose writing	Expose Pre-processing techniques NN architectures
February-May	Data pre-processing Implementation	Trained models Performance results
May-June	Experiments Evaluation	Comparable results
June-July	Writing of the results and documentation	Thesis

C. Preliminary Bibliography

Overview of literature sources identified so far (sorted alphabetically by authors' names)

References

[1] Patricia Bota, Chen Wang, Ana Fred, and Hugo Plácido da Silva. A review, current challenges, and future possibilities on emotion recognition using machine learning and physiological signals. *IEEE Access*, PP:1–1, 09 2019.

[2] Jun-Ho Choi and Jong-Seok Lee. Embracenet: A robust deep learning architecture for multimodal classification. *Information Fusion*, 51:259–270, 2019.

[3] Maciej Dziezyc, Martin Gjoreski, Przemysław Kazienko, Stanisław Saganowski, and Matjaz Gams. Can we ditch feature engineering? end-to-end deep learning for affect recognition from physiological sensor data. *Sensors*, 20:6535, 11 2020.

[4] F Eyben, M. Wöllmer, T. Poitschke, B. Schuller, C. Blaschke, B. Färber, and N. Nguyen-Thien. Emotion on the road—necessity, acceptance, and feasibility of affective computing in the car. *Advances in human-computer interaction*, 2010.

[5] M. Gjoreski, M. Ž. Gams, M. Luštrek, P. Genc, J. Garbas, and T. Hassan. Machine learning and end-to-end deep learning for monitoring driver distractions from physiological and visual signals. *IEEE Access*, 8:70590–70603, 2020.

[6] M. Grimm, K. Kroschel, H. Harris, C. Nass, B. Schuller, G. Rigoll, and T. Moosmayr. On the necessity and feasibility of detecting a driver's emotional state while driving. *in Proceedings of the 2nd International Conference on Affective Computing and Intelligent Interaction*, 4738:126 – 138, 2007. Springer, Lisbon, Portugal.

[7] J. A. Healey and R. W. Picard. Detecting stress during real-world driving tasks using physiological sensors. *IEEE Transactions on Intelligent Transportation Systems*, 6(2):156–166, 2005.

[8] Hassan Ismail Fawaz, Germain Forestier, Jonathan Weber, Lhassane Idoumghar, and Pierre-Alain Muller. Deep learning for time series classification: a review. *Data Mining and Knowledge Discovery*, 33(4):917–963, 2019.

[9] J. Izquierdo-Reyes, R. Ramirez-Mendoza, M. Bustamante-Bello, J. Pons-Rovira, and J. Gonzales-Vargas. Emotion recognition for semi-autonomous vehicles framework. *International Journal on Interactive Design and Manufacturing (IJIDeM)*, 12(1):1447–1454, 2018.

[10] Eiman Kanjo, Eman M.G. Younis, and Chee Siang Ang. Deep learning analysis of mobile physiological, environmental and location sensor data for emotion detection. *Information Fusion*, 49:46 – 56, 2019.

[11] Mahdi Khezri, Mohammad Firoozabadi, and Ahmad Reza Sharafat. Reliable emotion recognition system based on dynamic adaptive fusion of forehead biopotentials and physiological signals. *Computer methods and programs in biomedicine*, 122, 07 2015.

[12] Chao Li, Zhongtian Bao, Linhao Li, and Ziping Zhao. Exploring temporal representations by leveraging attention-based bidirectional lstm-rnns for multi-modal emotion recognition. *Information Processing Management*, 57(3):102185, 2020.

[13] Florian Lingenfelser, Johannes Wagner, and Elisabeth Andre. A systematic discussion of fusion techniques for multi-modal affect recognition tasks. pages 19–26, 11 2011.

[14] Catherine Marechal, Dariusz Mikołajewski, Krzysztof Tyburek, Piotr Prokopowicz, Lamine Bougueroua, Corinne Ancourt, and Katarzyna Wegrzyn-Wolska. *Survey on AI-Based Multimodal Methods for Emotion Detection*, pages 307–324. Springer International Publishing, 2019.

[15] Yagya Raj Pandeya and Joonwhoan Lee. Deep learning-based late fusion of multimodal information for emotion classification of music video. *Multimedia Tools and Applications*, 80(2):2887–2905, 2021.

[16] Jie-Lin Qiu, Xiao-Yu Li, and Kai Hu. Correlated attention networks for multimodal emotion recognition. pages 2656–2660, 12 2018.

[17] Ahmad Sohaib, Shahnawaz Qureshi, Johan Hagelbäck, Olle Hilborn, and Petar Jerčić. Evaluating classifiers for emotion recognition using eeg. *Found. Augment. Cognit. Lecture Notes Comput. Sci.*, 8027:492–501, 07 2013.

[18] Tongshuai Song, Guanming Lu, and Jingjie Yan. Emotion recognition based on physiological signals using convolution neural networks. pages 161–165, 02 2020.

[19] S. Steidl, A. Batliner, B. Schuller, and D. Seppi. The hinterland of emotions: Facing the open-microphone challenge. In *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*, pages 1–8, Sep. 2009.

[20] Ramanathan Subramanian, Julia Wache, Mojtaba Khomami Abadi, Radu Vieriu, Stefan Winkler, and Nicu Sebe. Ascertain: Emotion and personality recognition using commercial sensors. *IEEE Transactions on Affective Computing*, PP:1–1, 11 2016.

[21] Salah Taamneh, Panagiotis Tsiamyrtzis, Malcolm Dcosta, Pradeep Buddharaju, Ashik Khatri, Michael Manser, Thomas Ferris, Robert Wunderlich, and Ioannis Pavlidis. A multimodal dataset for various forms of distracted driving. *Scientific Data*, 4(1):170110, 2017.

[22] P. Tzirakis, G. Trigeorgis, M. A. Nicolaou, B. W. Schuller, and S. Zafeiriou. End-to-end multimodal emotion recognition using deep neural networks. *IEEE Journal of Selected Topics in Signal Processing*, 11(8):1301–1309, 2017.

[23] Jingyuan Wang, Ze Wang, Jianfeng Li, and Junjie Wu. Multilevel wavelet decomposition network for interpretable time series analysis. pages 2437–2446, 07 2018.

[24] Zhiguang Wang, Weizhong Yan, and Tim Oates. Time series classification from scratch with deep neural networks: A strong baseline. pages 1578–1585, 05 2017.

[25] Zhong Yin, Mengyuan Zhao, Yongxiong Wang, Jingdong Yang, and Jianhua Zhang. Recognition of emotions using multimodal physiological signals and an ensemble deep learning model. *Computer Methods and Programs in Biomedicine*, 140:93–110, 03 2017.