POLITECNICO
MILANO 1863

Dipartimento di Elettronica, Informazione e Bioingegneria

Master Degree in Computer Science and Engineering

Thesis Title

and its subtitle

by:
Gioele Pozzi

matr.:
10454628

Supervisor:

Co-supervisor:
Clara Borrelli

Academic Year
2019-2020

POLITECNICO
MILANO 1863

Dipartimento di Elettronica, Informazione e Bioingegneria

Master Degree in Computer Science and Engineering

Titolo Tesi

e sottotitolo

Candidato:
Gioele Pozzi

matricola:
10454628

Relatore:

Co-relatore:
Clara Borrelli

Anno Accademico
2019-2020

Abstract

One of the most attractive functions of music is that it can convey emotion and modulate a listener's mood [1]. Music can bring to tears, console us when we are grieving and drive us to love.
Most important thing is that music information behavior studies have identified emotion as an important criterion used by people in music searching and organization. Now become important the field of music emotion recognition.

Sommario

Piacere, so Mario

Acknowledgements

This thesis is the result of almost a year of work at the Image and Sound Processing Lab. First I would thank my supervisor...

Thanks to friends.

Thanks to family.

N.S.

Contents

List of Figures

List of Tables

1 Introduction

1.1 Motivation

Music has an important role in human life. More important, is that music in capable to evoke different emotions for people, but how is structured the relationship between music and emotion? We don't know yet. It's a hard problem, which have very different fields of background, from computer science, machine learning and psychology.

Emotion-aware Music Information Retrieval has been difficult due to the subjectivity and temporal of emotion responses to music. The role of physiological signals related to emotions could potentially be exploited in emotion-aware music discovery.

Music is the vehicle for emotions, feelings, passion and actions. With the music the composer create a narration which is purely emotional.

Can we measure emotions related to music?

1.2 Outline of the thesis

This thesis is organized as follows:

After a brief introduction about the objective of the thesis, in chapter 2 and 3 is presented a complete overview about the main arguments in chapter 2, as Music Information Retrieval (MIR) and Music Emotion Recognition (MER), Electrodermal Activity (EDA) and other physiological data using on-body sensors.

Chapter 4 is devoted to a complete overview of the state of the art about the main aspects related to chapters 2 and 3 of this thesis, in order

to have a general idea about what has been done in the past and which results they have achieved.

In chapter 5 is presented how the dataset we have considered is structured and what results they have reached. Is also shown our implementation of the problem.

Chapter 6 is about the results we have achieved and the comparison between the PMEmo performances.

Finally Chapter 7, draws the conclusions and outlines possible future research directions.

1.3 Application fields

The work proposed in this thesis finds potential application in several fields. Thanks to the work of PMEmo that created a large dataset containing emotion annotations and electrodermal activity signal, we have the possibility to study the relationship between music emotion and physiological signals.
Music Browsing can be an important field of application, because it helps in general in finding, generally in large datasets, what music user are looking for. For example one application could be to create a playlist based on the emotion that songs produce in each of us. Another important application is given by understanding the relationship between music and emotion, which is a well known relationship but hard to find structural connection between the two.

2

Theoretical Background on MIR and MER

This chapter introduces the readers to the main basics about Music Information Retrieval and Music Emotion Recognition.

2.1 Music Information Retrieval

Music information retrieval (MIR) is the interdisciplinary science of retrieving information from music. MIR is a small but growing field of research with many real-world applications. Those involved in MIR may have a background in musicology, psychoacoustics, psychology, academic music study, signal processing, informatics, machine learning, optical music recognition, computational intelligence or some combination of these.

MIR is being used by businesses and academics to categorize, manipulate and even create music.
A few application to MIR can be:

- Recommended systems: several already exist, but few are based upon MIR techniques, instead making use of similarity between users or laborious data compilation as in Pandora[1].
- Intelligent and adaptive digital audio effects: aim of design a system that determine the settings of audio effects based on the audio content.

[1] https://www.pandora.com

- Track separation and instrument recognition: like extracting the original tracks as recorded, which could have more than one instrument played per track. Instrument recognition is about identifying the instruments involved into one track.
- Automatic music transcription: process of converting an audio recording into symbolic, such score or a MIDI file.
- Automatic categorization: common task of MIR is musical genre categorization and is the usual task for the yearly Music Information Retrieval Evaluation eXchange (MIREX).

2.2 Music Emotion Recognition

Music Emotion Recognition (MER) aim to research on modeling humans emotion perception of music [6], a research topic that emerges in the face of the explosive growth of digital music. Automatic MER allows users to retrive and organize their music collections in a fashion that is more content-centric than conventional metadata-based methods.
The main challenge is based on the human perception of emotions, their subjective nature of emotion perception. Building such a music emotion recognition system, however, is challenging be- cause of the subjective nature of emotion perception. One needs to deal with issues such as the reliability of ground truth data and the difficulty in evaluating the prediction result, which do not exist in other pattern recognition problems such as face recognition and speech recognition.
MER methods developed try to address the issues related to the ambiguity and granularity of emotion description, the heavy cognitive load of emotion annotation, subjectivity of emotion perception, and the semantic gap between low-level audio signal and high-level emotion perception.

2.2.1 Importance of Music Emotion Recognition

Music plays an important role in human life, even more in the digital age. Never before has such a large collection of music been created and accessed daily by people. Before with the use of compact audio formats with near CD quality such as MP3 and now on with the various streaming services, have greatly contributed to the tremendous growth of digital music libraries.

Conventionally, the management of music collections is based on catalog metadata, such as artist name, album name, and song title. As the amount of content continues to explode, this conventional approach may be no longer sufficient. The way that music information is organized and retrieved has to evolve to meet the ever increasing demand for easy and effective information access.

Music, is a complex acoustic and temporal structure, it is rich in

content and expressivity. When an individual engages with music as a composer, performer or listener, a very board range of mental processes is involved, including *representational* and *evaluative*. The representational process includes the perception of meter, rhythm, tonality, harmony, melody, form, and style, whereas the evaluative process includes the perception of preference, aesthetic experience, mood, and emotion. The term evaluative is used because such processes are typically both valences and subjective. Both the representational and the evaluative processes of music listening can be leveraged to enhance music retrieval. According to a study of Last.fm[2], emotion tagging is the third most frequent type of tags (first is genre and second locale) assigned to music pieces by online users.
Even if emotion-based music retrieval was a new idea, a survey conducted in 2004 from [7] showed that about 28.2% of the participants identified emotion as an important criterion in music seeking and organization.
The table 3.1 represent the responses of 427 subjects to the question *"When you search for music or music information, how likely are you to use the following search/browse options?"* [7].

Search/Browse by	Positive rate
Singer/Performer	96.2%
Title of work(s)	91.6%
Some words of the lyrics	74.0%
Music style/genre	62.7%
Reccomendations	62.2%
Similar artist(s)	59.3%
Similar music	54.2%
Associated usage	41.9%
Singing	34.8%
Theme(main subject)	33.4%
Popularity	31.0%
Mood/emotional state	**28.2%**
Time period	23.8%
Occasions to use	23.6%
Instrument(s)	20.8%
Place/event where heard	20.7%
Storyline of music	17.9%
Tempo	14.2%
Record label	11.7%
Publisher	6.0%

Table 2.1: Responses of 427 subjects to the question *"When you search for music or music information, how likely are you to use the following search/browse options?"*

[2]https://www.last.fm/home

Into another survey [8], they present findings from an exploratory questionnaire study featuring 141 music listeners (between 17 and 74 years of age) that offers some novel insights.
One of the most exciting but difficult endeavors in research on music is to understand how listeners respond to music. It has often been suggested that a great deal of the attraction of music comes from its "emotional powers". That is, people tend to value music because it expresses and induces emotions. The table 2.2 tries to resume the motivations to the answer *"Why do we listen to music?"*

Motive	Ratio
"To express, release and influence emotions"	47%
"To relax and settle down"	33%
"For enjoyment, fun, and pleasure"	22%
"As company and background sound"	16%
"Because it makes me feel good"	13%
"Because it's a basic need, I can't live without it"	12%
"Because I like, love music"	11%
"To get energized"	9%
"To evoke memories"	4%

Table 2.2: Responses of 141 subjects to the question *"Why do you listen to music?"*

Some music companies, like Allmusic.com[3], gives the possibility to search music by emotion labels. With these, the user can retrive and browse artists or albums by emotion.

Making computers capable of recognizing the emotion of music also enhances the way humans and computers interact. It is possible to play back music that matches the users mood detected from physiological, prosodic, or facial cues. A cellular phone equipped with automatic music emotion recognition (MER) function can then play a song best suited to the emotional state of the user; a smart space (e.g., restaurant, conference room, residence) can play background music best suited the people inside it.

2.2.2 Recognizing the perceived emotion of music

There is a relationship between music and emotions, that has been the subject of much discussion and research in many different disciplines, like philosophy, musicology, sociology.
In psychological studies, emotion are often divided into three categories:

- *Expressed emotion*: the ones the performer tries to communicate with the listener.

[3]https://www.allmusic.com/moods

- ***Perceived emotion***: represented by music and perceived by the listener.
- *Felt or Evoked emotion*: induced by music and felt by the listener.

MER focus on perceived emotions because they are less subjective than felt emotions and are often easier to conceptualize. This because felt emotions depends on personal factors and the situation in which the listener processes the song. From an engineering point of view, one of the main interests is to develop a computational model of music emotion and to facilitate emotion-based music retrieval and organization. MIR community has made many efforts for automatic recognition of the perceived emotion of music, various implementations will be presented further in chapter 4.
A typical approach to MER categorizes emotions into a number of classes and applies Machine Learning (ML) techniques to train a classifies. Usually are extracted some features of music to represent the acoustic property of a music piece. Typically, a subjective test is conducted to collect the ground truth needed for training the computational model of emotion prediction. Subjects are asked to report their emotion perceptions of the music pieces.
To learn the relationship between music features and emotion labels have been applied, such as Support Vector Machines (SVMs), Gaussian Mixture Models (GMMs), Neural Networks (NN) and k-nearest neighbor.
After training, the automatic model can be applied to classify the emotion of an input music piece, for example a schematic diagram of the *categorical approach* to MER can be seen in figure 2.1.

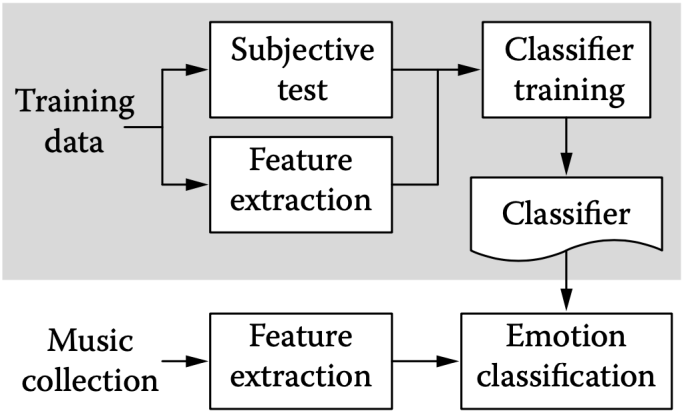

Figure 2.1: Schematic diagram of the categorical approach to MER

2.2.3 Open issues of Music Emotion Recognition

As MER is a quite new domain, there are some elements that have no clear answer. Four of these issues are:

1. Ambiguity and Granularity of emotion description: issue related to the relationship between emotions and the affective terms that denote emotions and the problem of choosing which and how many affective terms to be included in the taxonomy. Emotions are fuzzy concepts, there are main synonyms and similarities between different terms. In general, classification accuracy of an automatic model is inversely proportional to the number of classes considered [9].

2. Heavy cognitive load of emotion annotation: to collect data for training an automatic model, is typically conducted a subjective test by inviting human subjects to annotate the emotion of music pieces. The problem is that to reduce administrative effort, each music piece is annotated by two or three musical *experts* to gain consensus of the annotation result. Everyday contexts in which musical experts experience is so different from those non-experts require separate treatment. Since MER system is expected to be used in the everyday context, the emotion annotation should be carried out by *ordinary people*.

3. Subjectivity of emotional perception: music perception is intrinsically subjective and is under the influence of many factors such as cultural background, age, gender, personality and so forth. Therefore conventional categorical approaches that simply assign one emotion class to each music piece in a deterministic manner do not perform very well in practice.

4. Semantic gap between Low-Level (LL) and audio signal and High Level (HL) Human perception: it is difficult to accurately compute emotion values, and what intrinsic element of music causes a listener to create a specific emotional perception is still far from well understood.

2.2.4 Emotion description

Many researchers have suggested that music is an excellent medium for studying emotion, because people tend to make judgments about music and their affective responses to music.
Music represent emotions that are perceived by the listener or induced emotions that are felt by the listener. Now we will focus on the emotion conceptualization alone, since it's central to have a theoretical background to apply then to MER.
The celebrated paper of Hevner [10] , studied the relationship between music and emotions though experiments where subjects are asked to report some adjectives that came to their mind as the most representative part of a music played. From this have been proposed a large variety of emotion models, like the one presented and used in this thesis.

The idea of emotion conceptualization is to divide in two different approaches, the **Categorical approach** and the **Dimensional approach**.

Categorical approach

The first assumption of this emotion conceptualization is that emotions are categorized and categories are distinct from each other. For this approach, there is the idea that there are a limited number of innate and universal emotion categories such as:

- Happiness
- Sadness
- Anger
- Fear
- Disgust
- Surprise

All other emotions can be derived from these "*basic emotions*".
In psychological studies, different researchers have come up with different sets of basic emotions.
For example, another famous categorical approach to emotion conceptualization is Hevner's adjective checklist. He found eight clusters positioned in circle as in figure 2.2. The adjective within a cluster are similar, neighbor clusters varies in a cumulative way until reaching the opposite position where there is the contrast cluster. Hevner's checklist

Figure 2.2: Eight clusters proposed by Hevner

proposed in 1935 was suddenly updated and regrouped into ten groups by Fansworth and into nine groups in 2003 by Schubert.

Drawbacks of categorical approach is that the number of primary emotion classes is very small in comparison with the richness of music

emotion perceived by humans. The problem is in the sense that using a finer granularity, does not necessarily solve the problem because the language for describing emotions is inherently ambiguous and varies from person to person. Using a large number of emotion classes could submerge the subject and is impractical for psychological studies falsing results.

Dimensional approach

Categorical approach focuses mainly on the characteristics that distinguish emotions from one another, dimensional approach focuses on identifying emotions based on their position on a small number of emotion "dimensions" called axes, intended to correspond to internal human representation of emotion. These internal emotion dimensions are found by analyzing the correlation between affective terms.

There are several different names from past researchers gave very similar interpretations of the resulting factors like tension/energy, intensity/-softness, tension/relaxation for example. Most of the factors correspond to the two dimensions of emotion the *valence* (positive and negative affective states) and *arousal* (energy and stimulation level).
Russel, proposed a circumplex model of emotion in [11] which consist in a two-dimensional, circular structure as in figure 2.3 involving the dimensions of valence and arousal. In this structure, emotions that are inversely correlated, are placed across the circle from one another.

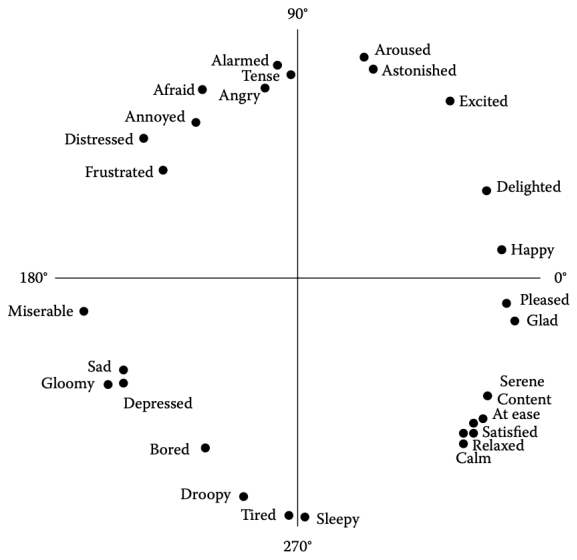

Figure 2.3: Russel's circumplex model of affect

Emotions that are easy to be confused, such as calm and sadness, appear to have similar valence and arousal values. This result implies that valence and arousal may be the most fundamental and most clearly communicated emotion dimensions among others. Also dimensional approach have its throwbacks, it is argued that dimensional approach blurs

important psychological distinctions and consequently obscure important aspects of the emotion process. One example in support of this argumentation is that anger and fear are placed close in the valence-arousal plane but they have very different implications for the organism. Also, it has been argued that using only a few emotion dimension cannot describe all the emotions without residuum.

Some researches, to overcome to these problems, tries to add a third dimension, called *potency* as dominant/submissive, to obtain a more complete picture of emotion. However, this would increase the cognitive load on the subjects at the same time, requires a more complex interface and makes hard to annotate the process. The third dimension problem is still in discussion.

Music Emotion Variation Detection

An important aspect that is not addressed in the previous two paragraphs is the temporal dynamics. Most researches has focused on music piece that are homogeneous with respect to the emotional plane. However, music can change its emotional expression during the song, becomes important to investigate the time-varying relationship between music and emotion. Here is more useful the dimensional approach to capture the continuous changes of emotional expression. Usually subjects are asked to rate valence and arousal in response of the stimulus every second.
For example, songs can be described by valence and arousal curves as in the following figure:

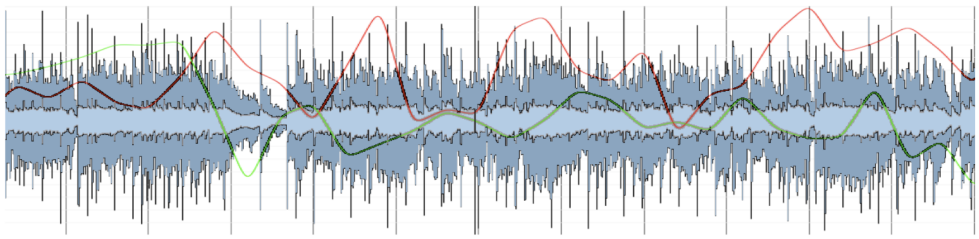

Figure 2.4: Valence and arousal curves for MEVD

2.2.5 Emotion recognition

MIR researches have been made to automate MER tasks, and the type of music under study has gradually shifted over the past few years from symbolic music to raw audio signal, from Western classical music to popular music. The purpose of MER is to facilitate music retrieval and management in the everyday music listening.

Nowdays are applied several machine learning techniques to recognize emotion from the music, and the training and automatic recognition model typically consists of the following steps:

1. Extract a certain number of features from audio signals to represent the music signal.

2. Collect from human annotators the ground truth emotion labels or emotion values.

3. Apply a learning algorithm between music features and emotion labels/values.

4. Predict emotion of an input song from the resulting computational model.

The music emotion recognition process can be schematized in the figure from [12]:

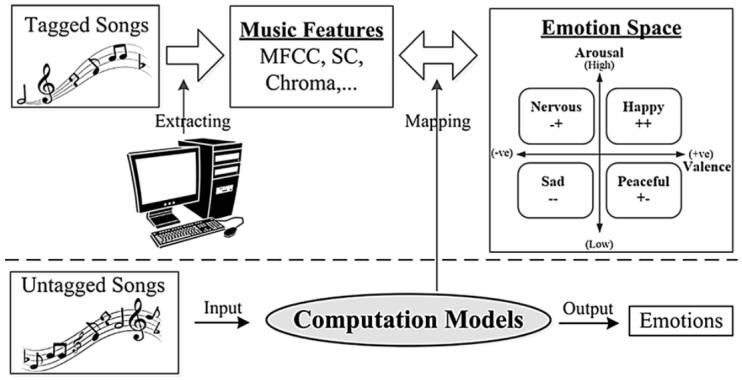

Figure 2.5: MER process

Researches that work on MER can be classified into three approaches.

The **categorical approach** that categorizes emotions into a number of discrete classes and applies machine learning techniques to train a classifier. The predicted emotion labels can be incorporated into a text-based or metadata-based music retrieval system.

The **dimensional approach** to MER defines emotions as numerical values over a number of emotion dimensions (valence and arousal). A regression model is trained to predict the emotion values that represent the affective content of a song, thereby representing the song as a point in an emotion space. Users can then organize, browse, and retrieve music pieces in the emotion space, which provides a simple means for user interface.

Categorical approach

Advantage of categorical approach is that it is easy to be incorporated into a text-based or metadata-based retrieval system. Emotion labels provide an atomic description of music that allows users to retrieve music through a few keywords. Here are present the issues discussed in chapter 2.2.3. The commonly adopted methods follows these points:

1. Data collection: nowadays there are several large-scale dataset covering all sort of music types and genres. Otherwise is desirable to collect data of the different types, getting rid of the effects called "*album effect*" or "*artist effect*" and collect a variety of music pieces. One problem is that there is no consensus on which emotion model or how many emotion categories should be used. Comparing systems that use different emotion categories and different dataset is impossible. However the issue concerning how many and which emotion classes should be used seem to remain open.

2. Data preprocessing: to compare music pieces fairly, music pieces are normally converted to a standard format, and since a complete music piece can contain sections with different emotions,a 20 to 30 second segment is often selected, which is representative of the song (like the chorus part). A good remark of the segment length can be found in [13].

3. Subjective test: emotion is a subjective matter, so the collection of the ground truth data should be conducted carefully. Annotation methods can be grouped into two categories:

   - Expert-based method: which employs a few musical experts to annotate emotions.
   - Subject-based method: employs a large number of untrained subjects to annotate emotions.

   The ground truth is set by averaging the opinion of all subjects (typically more than 10 subjects per song).
   It became important to not make a long test, in order to not compromise the reliability of the emotion annotations. Nowadays is introduced the use of listening games.

4. Features extraction: a certain number of features are extracted from the music signal to represent the different dimension of music listening like melody, timbre and rhythm.
   After features extraction, is applied feature normalization, in order to

5. Model training: the following step is to train a Machine Learning (ML) model to learn the relationship between emotion and music. Music emotion classification is carried out with classification ML algorithms, such as Neural Network, k-nearest neighbor (kNN), decision tree, Support Vector Machine (SVM) and Support Vector Classification (SVC).

Dimensional approach

The attractive part of dimensional approach is the valence-arousal

plane and the associated emotion-based retrieval methods. Due to the fact that the emotion plane contain an infinite number of emotion descriptions, the granularity and ambiguity issues are relieved.
Dimensional perspective is adopted to track the emotion variation of a classical song. The idea of representing the overall emotion of a popular song as a point in the emotion plane for music retrieval, under the assumption that the dominant emotion of a popular song undergoes less changes than a classical song. MER problem became a regression problem, and two independent models, called regressors, are trained to predict the valence-arousal values.
The dimensional approach requires the subjects to annotate the numerical valence-arousal values. This requirement impose an high cognitive load on the subjects.

Pros and cons of categorical and dimensional approach are schematized in the following table:

	Pros	Cons
Categorical	Intuitive Natural language Atomic description	Lack a unifying model Ambiguous Subjective Difficult to offer fine-grained differentiation
Dimensional	Focus on a few dimensions Good user interface	Less intuitive Semantic loss in projection Difficult to obtain ground truth

Table 2.3: Pros and cons of categorical and dimensional approaches

2.2.6 Valence and Arousal

As already mentioned before, the valence-arousal plane is the most used dimension plane to represent emotion.
In general, emotional experiences can be described by these two terms, *valence* (positive or negative affectivity) and *arousal* (calming or exciting). Some studies found that valence as well as intensity, is triggered by the amygdala, while the arousal by the reptilian brain.
The common framework for dealing with emotional experience is characterized in a two-dimensional space. Valence ranges from highly negative to highly positive, and arousal ranges from calming/soothing to exciting/agitating: High arousal emotional events are encoded better that non-arousing events. Instead of increasing overall attention to an event, an emotionally arousing stimulus decreased attentional resources available for information processing and focused attention only on the

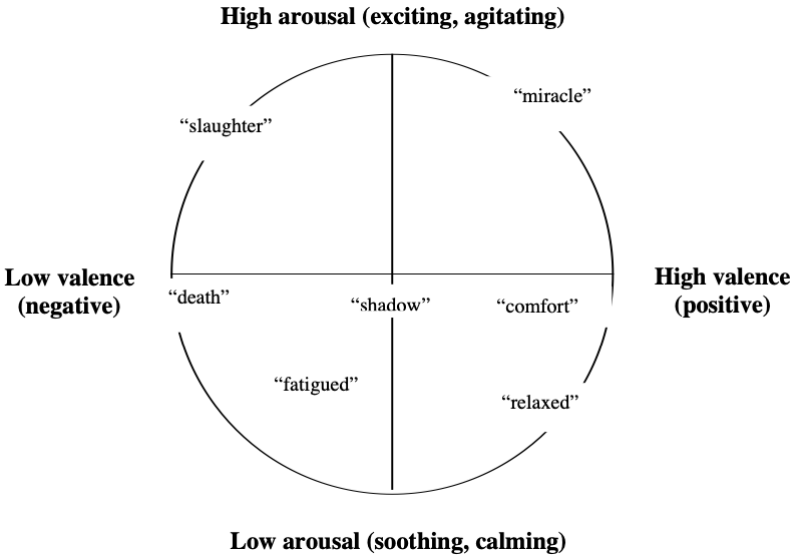

Figure 2.6: Valence and arousal plane, described in [2]

arousal-eliciting stimulus.

The experience of music listening is multidimensional. Different emotions are associated with different music patterns. For example, arousal is associated to:

- tempo (fast/slow)
- pitch (high/low)
- loudness (high/low)
- timbre (bright/soft)

while valence is associated to:

- mode (major/minor)
- harmony (consonant/dissonant)

as expressed in [14]. Emotion perception is correlated to the combination of music factor, rarely from just one of them. For example, loud chords and high-pitched chords tends to be feel as more positive valence than soft chords and low-pitched chords.

2.2.7 Music features

In MER analysis, an important step to define audio signals is to extract audio features and than apply a feature selection method.
There are several features that can be extracted from audio signal in order to represent five of the most useful perceptual dimensions of music listening:

- Energy: dynamic loudness, audio power, total loudness, specific loudness sensation coefficients.
- Rhythm: beat histogram, rhythm pattern, rhythm regularity, rhythm clarity, average onset frequency, average tempo.
- Temporal: zero-crossing, temporal centroid, log-attack-time.
- Spectrum: spectral centroid, spectral rolloff, spectral flux, spectral flatness.
- Harmony: salient pitch, chromagram centroid, harmonic change, pitch histogram.

These features are just an example of an infinite series of features that can be extracted from audio signals.

Gabrielsson et al. [14] noted that there are corresponding relations between the dimensional models and music features. Among these features, intensity is a basic feature, which is highly correlated with arousal and is used to classify the arousal dimension [15].
In [3] is shown a table summary of musical characteristics relevant to emotion, reported in 2.4. Despite the identification of these relations,

Features	Examples
Timing	Tempo, variation, duration, contrast
Dynamics	Overall level, crescendo/diminuendo, accents
Articulation	Overall staccato, legato, variability
Timbre	Spectral richness, harmonic richness
Pitch	High or low
Interval	Small or large
Melody	Range, direction
Tonality	Chromatic-atonal, key-oriented
Rhythm	Regular, irregular, smooth, firm, flowing, rough
Mode	Major or minor
Loudness	High or low
Musical form	Complexity, repetition, disruption
Vibrato	Extent, range, speed

Table 2.4: Musical features relevant to MER for [3]

many of them are not fully understood, still requiring further musicological and psychological studies, while others are difficult to extract from audio signals. Nevertheless, several computational audio features have been proposed over the years. While the number of existent audio features is high, many were developed to solve other problems (e.g., Mel-frequency cepstral coefficients (MFCCs) for speech recognition) and may not be directly relevant to MER.

Nowadays is not really clear the relationship between low-level and

mid-level features and mood. In order to capture different aspects is extracted a large set of features. This create a feature matrix that is then normalized in order to map them on the same range of values.

After the feature matrix is created is applied a feature selection or feature reduction algorithm to select the best set of features. Feature selection algorithms are based on two different ideas:

- High-level point of view: find the set of features that best model the concept. This lead to the accuracy of machine learning techniques being limited because of the limitation of the hypothesis done.
- Low level point of view: find the set of features that produces the best classification rate.

From the machine learning point of view, features are not necessarily of equal importance or quality, and irrelevant or redundant features may lead to inaccurate conclusion. Experiments have shown that, although the performance can thus be improved to a certain extent, using too many features leads to performance degradation [15].
With an highly discriminant sets of features, is not true that their combination produces a better discriminant power, for example if the set of features is 60, the number of possible combinations are:

$$n_{combinations} = \sum_{n=1}^{60} \binom{60}{k} \tag{2.1}$$

which is clearly impossible to compute, for this reason is applied some feature selection algorithms.

An example of feature selection for the categorical approach is the Sequential Feature Selection. It starts from an initial condition, and features are added or removed from a candidate subset while evaluating the *criterion* in two possibilities:

1. Sequential Forward Selection (SFS): features are sequentially added to an empty candidate set until the addition of further features does not decrease the criterion.
2. Sequential Backward Selection (SBS): features are sequentially removed from a full candidate set until the removal of further features increases the chosen criterion.

Another feature selection method is the Minimum-Redundancy-Maximum-Relevance (mRMR) which select the features with the highest relevance to the target class. Relevance is characterized in terms of *mutual information* which is defined as (given X and Y a pair of random variables):

$$I(x, y) = \iint p(x, y) \log \frac{p(x, y)}{p(x)p(y)} dxdy \tag{2.2}$$

where $p(x, y$ is the joint probability mass function of X and Y, $p(x)$ and $p(y)$ are the marginal probability mass function of X and Y respectively.

On the other side, for dimensional approach, feature selection is for example RReliefF [16]. Basic idea of this algorithm is that try to estimate the quality of each attribute (in this context the features) according to how well their values distinguish between instances that are close each other.
The pseudocode of the RReliefF feature selection algorithm from [16]:

```
INPUT: training data {x_i}_{i=1}^N, {y_i}_{i=1}^N, parameters K, σ, n
OUTPUT: vector W of estimations of the importance of features
  set N_dC, N_dM[m], N_dC&dM[m], W[m] to 0
  for t = 1 to n
    randomly select an instance i
    select k instances nearest to i
    for each neighbor j
      N_dC = N_dC + diff(y_i, y_j) · d(i, j)
      for m = 1 to M
        N_dM[m] = N_dM[m] + diff(x_im, x_jm) · d(i, j)
        N_dC&dM[m] = N_dC&dM[m] + diff(y_i, y_j) · diff(x_im, x_jm) · d(i, j)
      end
    end
  end
  for m = 1 to M
    W[m] = N_dC&dM[m]/N_dC − (N_dM[m] − N_dC&dM[m])/(n − N_dC)
  end
```

Figure 2.7: RReliefF pseudocode

Another feature selection for dimensional approach is Principal Component Analysis (PCA) and Independent Component Analysis (ICA). The method starts with all features and reduces them one by one, and hence is similar to backward selection. The goal of ICA is to find a linear representation of non-Gaussian data so that the components are statistically independent, or as independent as possible. While the other well known linear transformation methods (PCA) benefit from the gaussianity of the data, ICA improves the classifier performance in the opposite case.

2.2.8 Machine learning

Machine learning (ML) is the scientific study of algorithms and statistical models that computer systems use to perform a specific task without using explicit instructions, relying on patterns and inference instead. It is seen as a subset of artificial intelligence (AI). Machine learning algorithms build a mathematical model based on sample data, known as "training data", in order to make predictions or decisions without being explicitly programmed to perform the task.

When most people hear about ML they picture a robot, but it not just fantasy, it's already here, it has been around for decades in some specialized applications like *Optical Character Recognition*. The first ML application that became mainstream was done in 1990s, the *spam filter* [17].
A classical definition came from *Arthur Samuel* in 1959:

"*Machine Learning is the field of study that gives computers the ability to learn without being explicitly programmed*"

Another definition, more engineering-oriented is by *Tom Mitchell* in 1997:

"*A computer program is said to learn form experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E*"

The main difference between traditional programming and ML is well schematized in the figure 2.8 There are many different Machine Learning

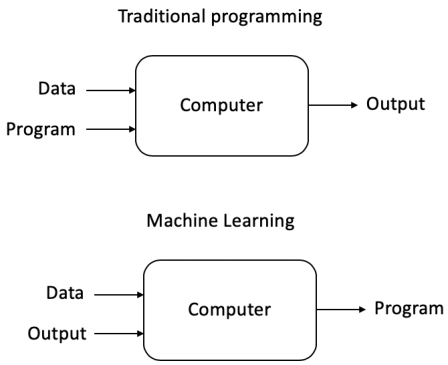

Figure 2.8: Traditional programming versus Machine Learning

systems. They can be classified in categories based on:

- Whether or not they are trained with human supervision (supervised, unsupervised, reinforcement learning).
- Whether or not they can learn incrementally on the fly (online and bach learning).
- Whether they work by comparing new data points to know data points, or instead detect patterns in the training data and build a predictive model (instance-based and model-based learning).

These criteria are not exclusive, they can be combined together.

In supervised learning, the algorithm builds a mathematical model from a set of data that contains both the inputs and the desired outputs. For example, if the task were determining whether an image contained a certain object, the training data for a supervised learning algorithm would include images with and without that object (the input), and each

image would have a label (the output) designating whether it contained the object.
Semi-supervised learning algorithms develop mathematical models from incomplete training data, where a portion of the sample input doesn't have labels.
In unsupervised learning, the algorithm builds a mathematical model from a set of data that contains only inputs and no desired output labels.

Classification algorithms and regression algorithms are types of supervised learning. Classification algorithms are used when the outputs are restricted to a limited set of values. For a classification algorithm that filters emails, the input would be an incoming email, and the output would be the name of the folder in which to file the email. For an algorithm that identifies spam emails, the output would be the prediction of either "spam" or "not spam", represented by the Boolean values true and false. Regression algorithms are named for their continuous outputs, meaning they may have any value within a range.
In this thesis the focus will be on **supervised learning**.

Supervised learning algorithms build a mathematical model of a set of data that contains both the inputs and the desired outputs. The data is known as **training data**, and consists of a set of training examples. Each training example has one or more inputs and the desired output, also known as a supervisory signal. In the mathematical model, each training example is represented by an array or vector, called **feature vector**, and the training data is represented by a **matrix**. Through iterative optimization of an objective function, supervised learning algorithms learn a function that can be used to predict the output associated with new inputs. An optimal function will allow the algorithm to correctly determine the output for inputs that were not a part of the training data. An algorithm that improves the accuracy of its outputs or predictions over time is said to have learned to perform that task.
In order to solve a problem of supervised learning, one has to perform steps:

1. Determine the training data type.

2. Gather a training set.

3. Determine the input feature representation of the learned function. Here input objects are transformed into feature vector which contains a number of features that describe the object.

4. Determine the structure of the learned function and corresponding algorithm.

5. Run the algorithm on the training set and optimize performances on a subset called *validation set* of the training set, or through *cross-validation*.

6. Evaluate the accuracy of the model.

There are several algorithms of supervised learning, there are no one that works best on all problems, due to this different algorithms are tested. Most widely used learning algorithms are:

- Support Vector Machines (SVM).
- Support Vector Regression (SVR).
- Linear Regression (LR).
- Decision Tree (DT).
- Neural Networks (NN).

Regression and Classification are both problems of supervised machine learning, the main difference between them is that the output variable in regression is numerical (or continuous) while that for classification is categorical (or discrete).

The task of MER is a regression problem both for dimensional, categorical and MEVD. In dimensional approach, the valence-arousal plane with a continuous space. Each point of the plane is considered an emotion state. This allow to overcome the categorical problem of granularity issue since the emotion plane implicitly offers an infinite number of emotion descriptions.
The regression approach applies a computational model that predicts the valence and arousal values of a music piece, which determine the placement of the music piece in the emotion plane [6].

A user can then retrieve music by specifying a point in the emotion plane according to his/her emotion state, and the system would return the music pieces whose locations are closest to the specified point. Because the 2D emotion plane provides a simple means for user interface, novel emotion-based music organization, browsing, and retrieval can be easily created for mobile devices.

Regression approach

A schematic diagram of the regression approach is in 2.9 where in the training phase, regression model are trained by learning the relationship between music features x and ground truth emotion values y. To denote regressors for valence and arousal are used r_V and r_A. In the test phase, given the features x_* of an input song, the regressors r_V and r_A can be applied to predict its emotion values $y_* = [v_*, a_*]^T = [r_V(x_*), r_A(x_*)]^T$. The regression theory aim at predicting a real value from observed variables, in MER application music features. The VA values are predicted directly from music features and due to this MER can be approached as a regression problem.

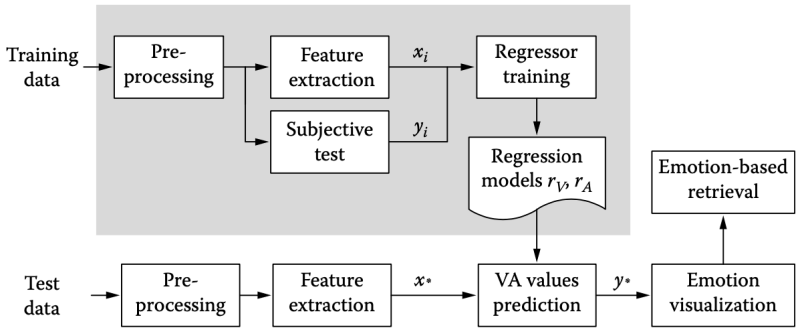

Figure 2.9: Schematic diagram of a regression approach

Given N inputs $(\mathbf{x}_i, y_i)$, with $i \in 1, ..., N$ where $\mathbf{x}_i$ is the feature vector of an object d_i (music piece), and y_i is the real value to be predicted (valence or arousal), a regressor $r(\cdot)$ is created by minimizing the mean squared error (MSE) ε:

$$\varepsilon = \frac{1}{N}\sum_{i=1}^{N}(y_i - r(\mathbf{x}_i))^2 \tag{2.3}$$

where $r(\mathbf{x}_i)$ is the prediction result for d_i.
In this thesis in mathematical expressions, **bold** font represent vectors and matrices.

To evaluate the performances of the regression approach with various ground truth data spaces, feature spaces and regression algorithms is used the R^2 statistics, which is a standard way for measuring the goodness of fit of regression models. It is calculated as:

$$R^2(\mathbf{y}, r(\mathbf{X})) = 1 - \frac{N\varepsilon}{\sum_{i=1}^{N}(y_i - \widehat{y})^2} = 1 - \frac{\sum_{i=1}^{N}(y_i - r(\mathbf{x}_i))^2}{\sum_{i=1}^{N}(y_i - \widehat{y})^2} \tag{2.4}$$

where $\widehat{y}$ is the mean of the ground truth. R^2 is comparable between experiments because of the normalization of the total squared error $N\varepsilon$ by the energy of the ground truth. The value of R^2 lies in $[-\inf; 1]$ where $R^2 = 1$ means the model perfectly fits the data, while a negative R^2 means the model is even worse than simply taking the sample mean.

The regression approach to MER, however, is not free of issues. First, the regression approach suffers from the subjectivity issue of emotion perception as it assigns the valence and arousal values to a music piece in a deterministic way. It is likely that different users perceive different emotion values in the music piece. Second, the regression approach requires numerical emotion ground truth to train the computational model, but performing such an emotion rating is a heavy cognitive load to the subjects.

3 Theoretical Background on EDA

This chapter introduces the readers to Electrodermal Activities and how are they related with Music Emotion Recognition task.

3.1 Electrodermal phenomena

Already in the 80's, psychological factors related to electrodermal phenomena were observed. It became an important field of study, due to the fact its ease of obtaining a distinct electrodermal response (EDR), the intensity of which seems apparently related to stimulus intensity and/or its psychological significance [18].

While there is still widespread disagreement and confusion about the nature and causes of musically evoked emotions, recent studies involving real-time observation of brain activity seem to show that areas of the brain linked with emotion (as well as pleasure and reward) are activated by music listening [19].

3.2 Electrodermal Activity

Electrodermal Activity (EDA) was first introduced by Johnson and Lubin in 1966 [20] as a common term for all electrical phenomena in skin, including all active and passive electrical properties that can be traced back to the skin and its appendages.
Electrodermal activity (EDA) is the property of the human body that causes continuous variation in the electrical characteristics of the skin. Historically, EDA has also been known as skin conductance, galvanic skin response (GSR), electrodermal response (EDR), psychogalvanic re-

flex (PGR), skin conductance response (SCR), sympathetic skin response (SSR) and skin conductance level (SCL). The long history of research into the active and passive electrical properties of the skin by a variety of disciplines has resulted in an excess of names, now standardized to electrodermal activity.

The use of the term *response* for phasic electrodermal phenomena suggests that there is a distinct relationship to a stimulus producing an EDR. Sometimes there are phasic parts that cannot be traced to any specific simulation, they are called *spontaneous* or *non-specific* EDRs.

There is ample empirical evidence that electrodermal phenomena are generated by sweat gland activity in conjunction with epidermal membrane processes. When sweat gland activity is abolished in humans, either as a result of congenital absence, by sympathectomy, by peripheral sudomotor nerve discharge, or by pharmacological blocking, SCRs and SPRs are normally eliminated and SCL is considerably reduced [21].

As mentioned before, skin conductance is characterized by:

- Tonic skin conductance level (SCL): smooth underlying slowly changing level, it accounts for the general levels of the coductivity of the skin.
- Phasic skin conductance response (SCR): rapidly changing peaks, results from momentary sympathetic activation when arousing stimuli are present.

In the figure 3.1 can be seen an EDA file plot and the division in Tonic and Phasic parts thanks to pyphysio library [22].

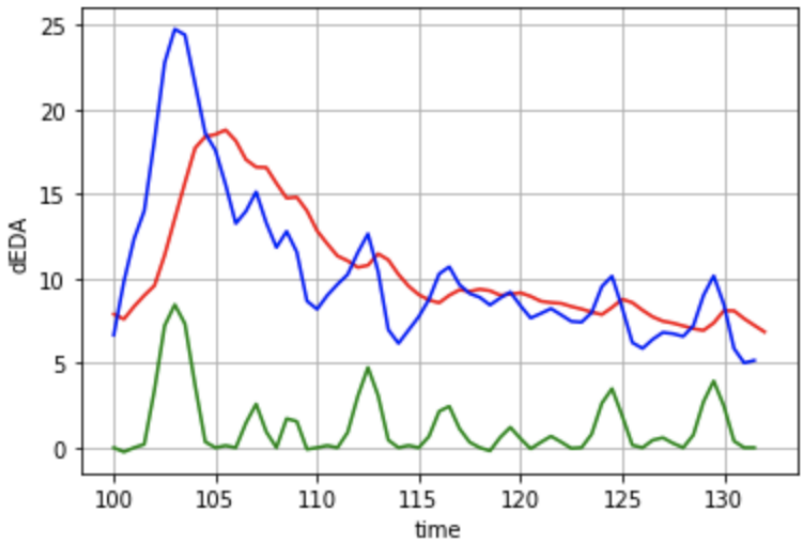

Figure 3.1: Representation of EDA signal (in red), driver signal (in blue) and phasic signal (in green)

Another useful graph is shown in the figure 3.2 from [23] that represent

division between the complete signal (in blue), the tonic component (in green-dashed) and the phasic component (in red).

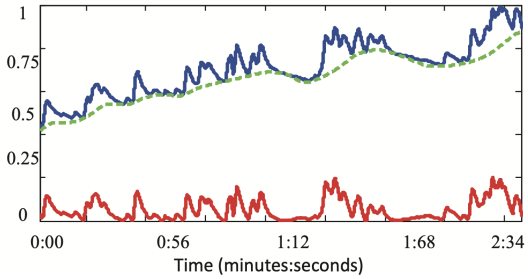

Figure 3.2: Representation of EDA signal (in red), driver signal (in blue) and phasic signal (in green)

The time series of the change of Skin Conductance (SC) can be characterized by a slowly varying tonic activity and fast varying phasic activity. The SCRs shows a steep incline to the peak and a slow decline to the basilne. The successions of SCRs usually results in a superposition of subsequent SCRs as one SCR arises on top of the declining trail of the preceding one.
The figure 3.3 from [24] shows a SC data section of 165 s, the upper row shows the original SC data. The middle row shows the driver signal which results from deconvolution of the SC data. Inter-impulse data are used to estimate the tonic part of the driver at 10-s intervals (tonic grid points). The tonic driver is used to compute the tonic SC (see upper row). Subtraction of the tonic part from the driver results in the phasic driver (lower row). The phasic driver shows a virtually zero baseline and distinct phasic responses.

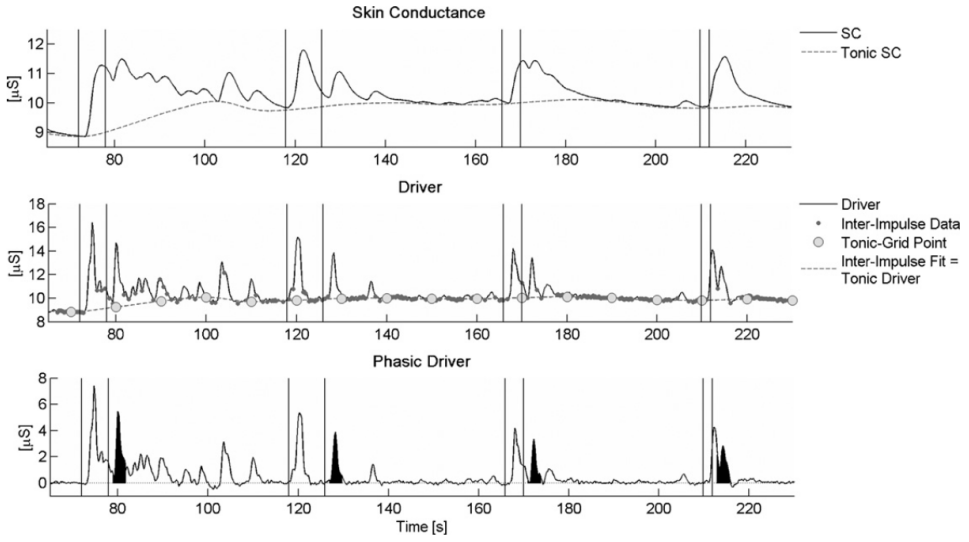

Figure 3.3: Skin conductance and phasic driver extraction

3.2.1 Measurement principles

EDA can be measured both without externally applied voltage (endosomatic method) or with application of Direct Current (DC) or Alternating Current (AC) (exosomatic method). The widespread used method is the exosomatic with DC recordings. With direct voltage, skin resistance measurements will result when current is constant, while skin conductance measurement will result when voltage is kept constant.

There are some factors that should be controlled as possible sources or variance in EDA recordings, like environmental conditions as the climatic conditions and physiological factors like age, gender and ethnic differences.

EDA can be measured in many different ways electrically including skin potential, resistance, conductance, admittance, and impedance. It achieves this by passing a minuscule amount of current between two electrodes in contact with the skin. The units of measurement for conductance are microSiemens (μS).

3.2.2 Recording techniques

Electrodermal recording is usually performed with two electrodes. Exosomatic techniques use two active sites, while endosomatic recording requires an active and an inactive site.
Figure 3.4 illustrate the preferred palmar recording areas for exosomatic and endosomatic EDA recordings. Sites A and B for bipolar recordings. C and D for volar electrode sites.

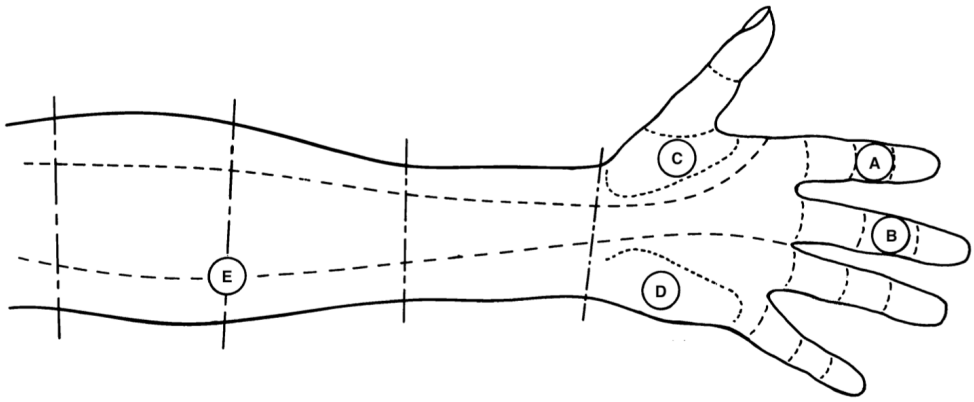

Figure 3.4: Preferred palmar recording areas for exosomatic and endosomatic EDA recordings

3.2.3 Artifacts identification in EDA data

EDA data is often captured by wearable devices, which makes the signal collected vulnerable to several types of noise. Artifacts can be generated from electronic noise or variation in the contact between the skin and

the recording electrode caused by pressure, excessive movement or adjustment of the device [25].
They may be mistaken for a skin conductance response, and this must be avoided.

Typically, as Boucsein [18] report, the shape of an SCR typically lasts between $1a$ to $5s$, has a steep onset and an exponentially decay and reaches an amplitude of at least $0.01\mu S$. An example of a typical SCR in figure 3.5.

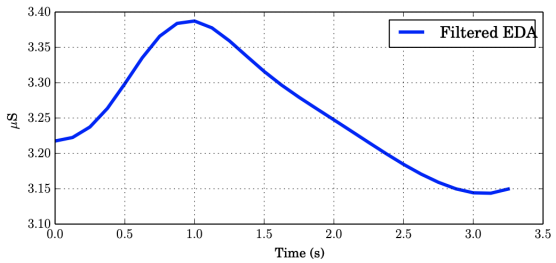

Figure 3.5: Example of a SCR shape

Currently, many researchers deal with signal artifacts and noise by applying exponential smoothing or low-pass filtering.
Additionally, filter cutoff frequencies are based only loosely on prior knowledge of typical characteristics of SCR shape, and vary widely study to study (1 and 5 Hz). The cutoff frequency ultimately chosen for a study is specific to that particular study, making generalization difficult.

There are much relevant techniques that are also able to recognize and compensate for large-magnitude artifacts that can result from pressure or movement of the device during recordings.
In [25] is presented a figure reported here 3.6, which shows a portion of signal that contains three artifacts, in which the fast decrease could not be produced by human physiology. Comparing the raw signal and the filtered version, the low-pass filter has not removed the artifacts.

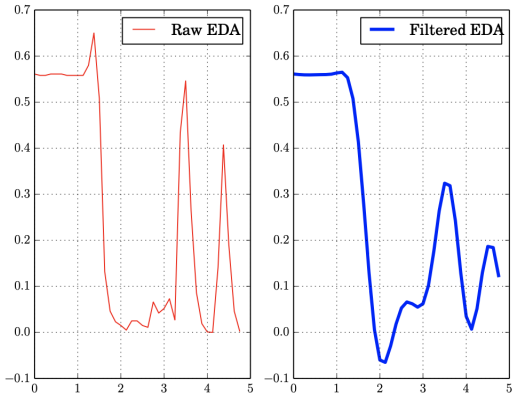

Figure 3.6: Portion of an EDA signal, the raw signal on the left in red, a 1 Hz low-pass filter applied on the signal to the left in blue

Some researchers, as Boucsein analysis [18], develop heuristic techniques for removing atypical portion of the EDA signal. Someone decide to discard portion of their data where the signal increased more than 20% per second or decreased more than 10% per second.
In another case, a study which collected EDA from two sensors (on both the ankle and wrist) [26] was able to detect artifacts by looking for epochs when only one of the two sensors had an abnormally low signal, or showed an unusually rapid increase or decrease.

In [25] developed a Machine Learning algorithm for automatically detecting EDA artifacts, providing empirical evaluation of classification performances.

3.2.4 EDA features

As for MER analysis, also in EDA data analysis, is important to find which features need to be extracted and than which feature selection method must be carried.

Emotion recognition from EDA has been commonly used for the assessment of user's experience in a variety of contexts such as recreational and games [27] and driving [28]. Previous research has explored the predictive power of a diverse set of EDA features of different types, including time domain, frequency domain, and time-frequency domain features.

Regarding time domain features, most usually features considered are the statistical parameters of the signal as:

- Mean value: μ is the central value of a discrete set of numbers $x_1, x_2, ..., x_n$, specifically, the sum of the values divided by the number of values:
$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{3.1}$$

- Standard deviation: is a measure of the amount of variation or dispersion of a set of values:
$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2} \tag{3.2}$$

- Kurtosis: is a measure of the "tailedness" of the probability distribution of a real-valued random variable:
$$kurt = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^4}{\sigma^4} \tag{3.3}$$

- Skewness: is a measure of the asymmetry of the probability distribution of a real-valued random variable about its mean:
$$skew = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^3}{\sigma^3} \tag{3.4}$$

In [4], as an example, were extracted the features for EDA data shown in table .

Domain	Feature vector	Number of features
Time	SCR related	7
	Statistical features	8
	Hjorth features	2
	Higher Order Crossing	5
Frequency	Statistical features	8
	Band power	9
Time-Frequency	DWT coefficients	56
	SWT features	40
	MFCCs	481
	Statistical features MFCC	5

Table 3.1: Features extracted in [4]

Other cases, researchers have focused on event-related features of EDA. They are useful when are presented to the subjects some events, stimulus, like images or sounds.
Examples of event-related aspects of EDA considered in other studies are SCR amplitude, SCR peak count, mean SCR rise time, or the sum of SCR areas.

Fewer researches, as [4] remarks, has focused on the predictive power of EDA related to the frequency domain. The frequency domain analysis has shown superior capability for the gradient component's detection of individual SCR.
Due to the different rate of physiological process, EDA signals vary significantly with the frequency [29].
Frequency oscillations of EDA signals can be divided into different frequency sub-bands to analyze it. Indeed previous researchers has considered statistical aspects.

As for audio, also for EDA data, after constructing a feature matrix, need to apply an algorithm of feature selection to improve data reliability.

Some examples of feature selection could be:

- Joint Mutual Information (JMI): focuses on the increasing complementary information between features.
- Conditional Mutual Information Maximization (CMIM): it can properly identify truly redundant features and noisy features, and gives preference to informative, uncorrelated features.
- Double Input Symmetrical Relevance (DISR): a normalized variant of JMI.

In general it is not known which features are most appropriate for emotion recognition from EDA and previous works have made limited contributions on a systematic comparison of EDA features.
In [4] there is a table showing various features extracted for EDA signals presented in literature.

4
State of the Art

This chapter introduces the readers to a complete review of the problem and all the different resolution possibilities.

4.1 Physiological signals

In this section will be defined a general overview on physiological signals that can be used in order to achieve a solution to the Music Emotion Recognition problem.

Emotions, which affect both human physiological and psychological status, play a very important role in human life. Positive emotions help improve human health and work efficiency, while negative emotions may cause health problems. Long term accumulations of negative emotions are predisposing factors for depression, which might lead to suicide in the worst cases.
The emotion often refers ti a mental state that arises spontaneously rather than through conscious effort and it is accompanied by physical and physiological changes, relevant to the human organs and tissues such as brain, heart, skin, blood flow, muscle, facial expressions, voice, etc. [30].

Emotion recognition has been applied in many areas such as safe driving [31], health care especially mental health monitoring [32], social security [33], and so on.
In general, emotion recognition methods could be classified into two major categories:

- Using human physical signals such as facial expression [34], speech [35], gesture, posture, etc. This method has the advantage of easy

collecting and is a chapter which has been studied for years. On the other side, the reliability cannot be guaranteed, as it is relatively easy for people to control the physical signals like facial expression or speech to hide real emotions, especially during social communications.

- Using internal signals as:
  - Electroencephalogram (EEG)
  - Electrocardiogram (ECG)
  - Electromyogram (EMG)
  - Blood Pressure (BP)
  - Hearth Rate Variability (HRV)
  - Electrodermal Activity (EDA) as:
    - Skin Resistance (SR)
    - Skin Temperature (ST)
    - Skin Conductivity (SC)
    - Galvanic Skin Response (GSR)
  - Respiration (RSP)

  These signals are produced by the Nervous System which is divided into:

  - Central Nervous System (CNS)
  - Peripheral Nervous System (PNS): consist of the autonomic and somatic nervous systems (ANS and SNS).

  EEG, ECG, EMG, GSR, RSP and GSR change in a certain way when people face some specific situations. Physiological signals are in response to the CNS and ANS. Due to the fact that CNS and ANS are involuntarily activated, they cannot be controlled.

In the table 4.1 is shown a summary of various papers using different biological signals.

Biological signal	Paper
ECG	[36], [37], [38], [39], [40]
ECG, EMG, RSP	[41]
ECG, GSR	[42]
HR, SR	[43]
EEG	[44]
HR	[45]

Table 4.1: Papers with correspondent biological signal used

In table 4.2 is presented the relationship between emotions and physiological features, thanks to [30]. Arrows indicate increased (↑), decreased (↓), no change in activation from the baseline (−) or both increases and decreases in different studies (↑↓).

Signal	Anger	Anxiety	Embarrassment	Fear	Amusement	Happiness	Joy
Cardiovascular							
HR	↑	↑	↑	↑	↑↓	↑	↑
HRV	↓	↓	↓	↓	↑	↓	↑
LF		↑		(−)		(−)	
LF/HF		↑			(−)		
PWA				↑			
PEP	↓		↓	↓	↑	↑	↑↓
SV	↑↓	(−)		↓		(−)	↓
CO	↑↓	↑	(−)	↑	↓	(−)	(−)
SBP	↑	↑	↑	↑	↑	↑	↑
DBP	↑	↑	↑	↑	↑	↑	(−)
MAP			↑	↑	↑	↑	
TPR	↑			↓	↑	↑	(−)
FPA	↓	↓		↓	↓	↑↓	
FPTT	↓	↓		↓		↑	
EPTT		↓		↓		↑	
FT	↓	↓		↓	(−)	↑	
Electrodermal							
SCR	↑	↑		↑	↑		
nSRR	↑	↑		↑	↑	↑	↑
SCL	↑	↑	↑	↑	↑	↑	(−)
Respiratory							
RR	↑	↑		↑	↑	↑	↑
Ti	↓	↓		↓	↓	↓	
Te	↓	↓		↓		↓	
Pi	↑			↑		↓	
Ti/Ttot				↑	↓		
Vt	↑↓	↓		↑↓	↑↓	↑↓	
Vi/Ti						↑	
Electroencephalography							
PSD(α wave)	↑	↑		↓	↑	↑	↑
PSD(β wave)	↓				↑		
PSD(γ wave)				↓	↑	↑	↑
DE (avg)	↑	(−)		↓		↑	↑
DASM (avg)	(−)			↑	↓	↓	↓
RASMs (avg)	↑			↑		↓	

Table 4.2: Relationship between emotions and physiological features

The position of different biosensors is shown in figure 4.1.

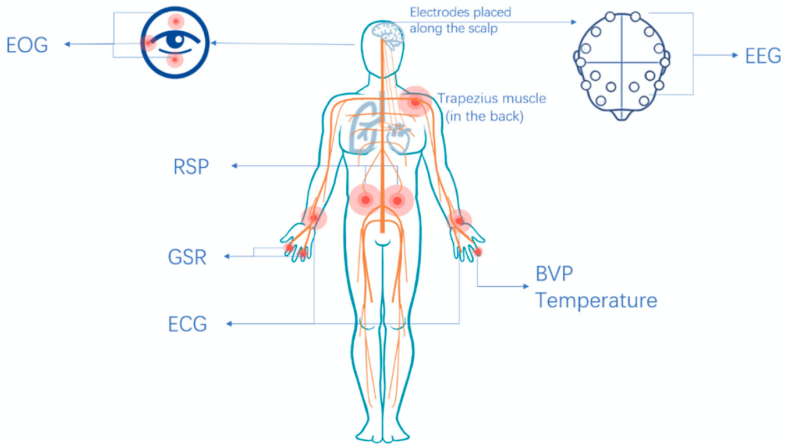

Figure 4.1: Position of the bio-sensors

4.1.1 Electroencephalogram

Electroencephalogram (EEG) is an electrophysiological monitoring method to record electrical activity of the brain. EEG measures voltage fluctuations resulting from ionic current within the neurons of the brain. Clinically, EEG refers to the recording of the brain's spontaneous electrical activity over a period of time, as recorded from multiple electrodes placed on the scalp.

EEG is most often used to diagnose epilepsy, which causes abnormalities in EEG readings.[2] It is also used to diagnose sleep disorders, depth of anesthesia, coma, encephalopathies, and brain death.
Many studies have indicated that the physiological correlates of emotions are likely to be found in the central nervous system rather than simply in peripheral physiological responses. Researchers have supported this viewpoint using EEG or other neuroimaging (e.g., functional Magnetic Resonance Imaging) approaches to investigate the specificity of brain activity associated with different emotional states.
However, most of the available studies on emotion-specific EEG response have focused on EEG characteristics at the single-electrode level, rather than at the level of EEG-based functional connectivity.

4.1.2 Electrocardiogram

Electrocardiogram (ECG) is a recording of the electrical activity of the hearth using electrodes placed on the skin. These electrodes detect small electrical changes that are a consequence of cardiac muscle depolarization followed by repolarization during each cardiac cycle, the heartbeat.

There are three main components to an ECG: the P wave, which

represents the depolarization of the atria; the QRS complex, which represents the depolarization of the ventricles; and the T wave, which represents the repolarization of the ventricles, as in figure 4.2.

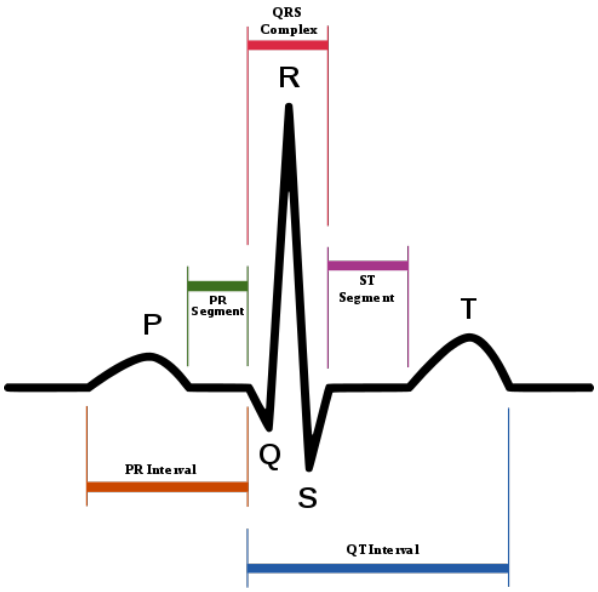

Figure 4.2: ECG of a heart in normal sinus rhythm

4.1.3 Electromyogram

Electromyogram (EMG) is an electrodiagnostic medicine technique for evaluating and recording the electrical activity produced by skeletal muscles.
An electromyograph detects the electric potential generated by muscle cells when these cells are electrically or neurologically activated. The signals can be analyzed to detect medical abnormalities, activation level, or recruitment order, or to analyze the biomechanics of human or animal movement.

Therefore, the best readings are obtained when the sensor is placed on the muscle belly and its positive and negative electrodes are parallel to the muscle fibers. Since the number of muscle fibers that are recruited during any given contraction depends on the force required to perform the movement, the intensity (amplitude) of the resulting electrical signal is proportional to the strength of contraction.

In psychophysiology, EMG was often used to find the correlation between cognitive emotion and physiological reactions. In the work by Sloan [46], for example, the EMG was positioned on the face (jaw) to distinguish *smile* and *frown* by measuring the activity of zygomatic major and corrugator supercilli. In experiment of [41], bipolar electrodes were placed at the upper trapezius muscle (near the neck) in order to measure the mental stress of the subjects.

4.1.4 Hearth Rate Variability

Hearth Rate Variability (HRV) measure the beat-to-beat temporal changes of the hearth rate, sometimes it is calculated from ECG, but the usability of measuring the ECG is limited. HRV can be evaluatd also through the Blood Volume Pulse (BVP) or Photoplethysmography (PPG).

A reduced HRV is linked to psychiatric illness as depression, anxiety. The heart rate is the most natural choice for arousal detection using comparison of sympathetic and parasympathetic frequency bands of the time series. However, it is highly dependent on the position of the body during monitoring.

4.1.5 Electrodermal Activity

As already been studied in chapter 3, EDA measures the resistance of the skin and the skin conductivity applying electrodes to the skin. The skin conductivity decreases during relaxed states, and increase when exposed to effort.

4.1.6 Respiration

Respiration (RSP) is the process of moving air into an out of the lungs to facilitate gas exchange with the internal environment, mostly bringing in oxygen and flushing out carbon dioxide.
The respiration can be measured with a latex rubber band, the amount of stretch in the elastic is measured as a voltage change and recorded. The most common measures of RSP are the depth of breathing and the rate of RSP.

RSP rate generally decreases with relaxation, tense situations may result in momentary RSP cessation. Irregularity in the RSP pattern could be the cause of negative emotions.
Due to the fact that RSP is closely linked to the cardiac function, RSP can be affect other measures like EMG and SC measurements.
Positions to the left, and typical waveform of the signals to the right are presented in the figure 4.3.

4.2 General methodology

For physiological signal-based emotion recognition, there is a common methodology which can be divided into two categories:

- Traditional Machine Learning methods: model specific methods, which require carefully designed hand-crafted features and feature optimization methods.

- Deep learning methods, model-free methods, which can learn the inherent principle of the data and extract features automatically.

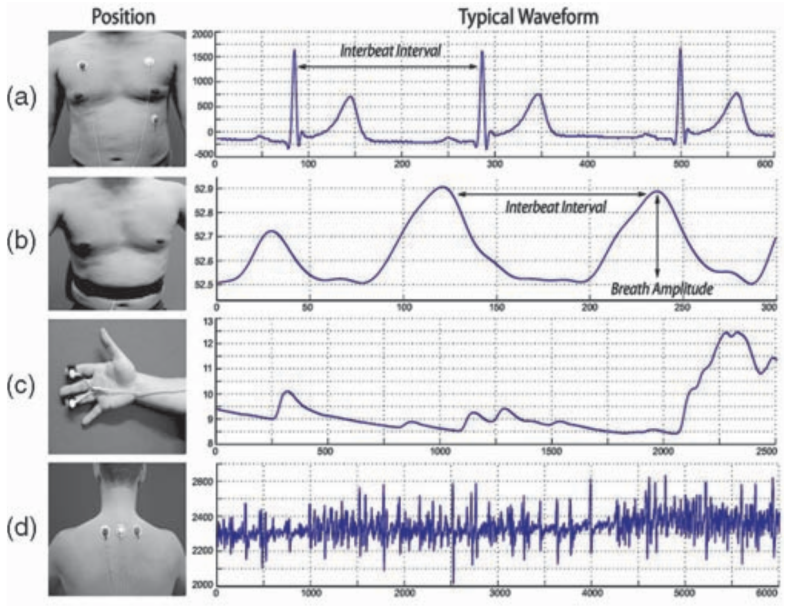

Figure 4.3: Positions (left) and waveform of the signals (right), (a) ECG, (b) RSP, (c) SC, (d) EMG

The whole emotion recognition framework is shown in figure 4.4.

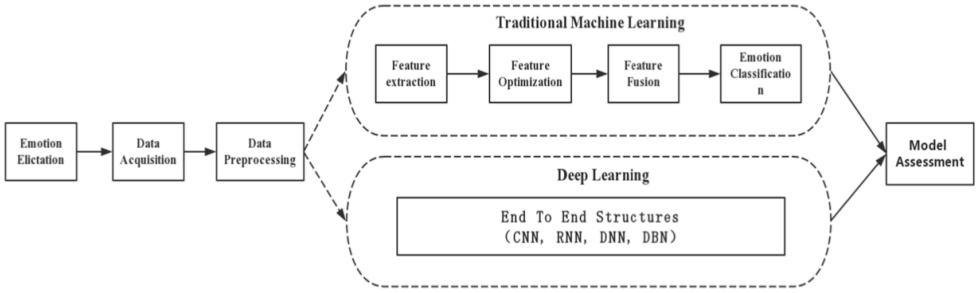

Figure 4.4: Emotion recognition process using physiological signals under targer emotion stimulation

4.2.1 Preprocessing

After the part of data acquisition, which is different for each physiological signal, there is the data preprocessing step. It is necessary to eliminate the noise effect, artifacts and other signal parts that may lead to wrong results. Due to the complex and subjective nature of raw physiological signals and the sensitivity to noises, electromagnetic interferences, movement artifacts, ... this step is mandatory.

Some of the common steps can be summarized in:

- Filtering: is commonly used a low-pass filter to remove noises, or also adaptive bandpass filters to remove artifacts.

- Discrete Wavelet Transform (DWT): used to reduce the noise of physiological signals
- Independent Component Analysis (ICA): used to extract and remove respiration sinus arrhythmia from ECG.
- Empirical Mode Decomposition (EMD): used to remove the eye-blink from EEG.

4.2.2 Traditional Machine Learning

Main steps for traditional ML methods, as already presented in 2.2.8. There are processes including feature extraction, feature selection and classification as reviewed in [47].

Feature extraction

Feature extraction plays a fundamental role in the emotion recognition model. Several major features are extracted for each physiological signal, because is important to extract the most prominent features for emotion recognition. For example EEG is a complex and non-stationary signal, so some statistical features like Power Spectral Density (PSD) and Spectral Entropy (SE) are commonly used.
Often are extracted statistical features as mean, standard deviation, Kurtosis, Skewness, entropy.
However, each bio-signal has to be investigated separately as extracted features might vary in their usefulness for the classification of emotions.

Feature selection

After the feature extraction process, there might be a quantity of features, some of which may be irrelevant, some that are probably correlated each other, there might be some redundant features.
It lead to a long time of analyzing the features and train the model and it is easy to produce overfitting problem and another problem of sparse features, called *curse of dimensionality*, which results in the decrease of the model performance.
Some of the main feature selection algorithms are RfeliefF, MRmR, Sequential Backward Selection (SBS) and Sequential Forward Selection (SFS), PCA, ...

In general there are several feature selection alforithms, some reduce the dimensionality by taking out some redundant or irrelevant featuresm other transform the original one into a new set of features. Performances of the feature selection algorithm depends on the classifier and the dataset, due to this, the perfect feature selection algorithm do not exist.

Classification

In emotion recognition, the major task is to assign the input signal to one of the given class sets. There are several classification models like Linear Discriminant Analysis (LDA), k-Nearest Neighbor (kNN), Support Vector Machine (SVM), Random Forest (RF), ...

4.2.3 Deep Learning

Deep Learning (DL) methods have the benefits to be model-free methods, so they do not depends on the specific model considered.
Examples of DL algorithms are Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), ... Neural network in general are particular types of ML methods which have as fundamental unit the node, loosely based on the biological neuron.

The relevant aspect of DL is that it can learn the inherent principle of the data and extract features automatically, so there is no need to extract feature and select the most relevant ones, which could lead to a better generalization of the problem.

4.2.4 Model assessment and selection

The generalization error of the classifier can be evaluated by experiments, where a testing set should be used to test the ability of the classifier to classify the new samples, and the testing error on the testing set could be viewed approximately as the generalization error.
The dataset is divided into two mutually exclusive sets, the training and the testing set. It is important to maintain the consistency of the data distribution as much as possible. In general, the experiment need to repeat several times with random division and then calculate the average value as the evaluation result.

There is also k-fold crossing-validation. the initial sampling is divided into K sub-samples. One sub-sample is used as the testing set and the other $K - 1$ samples are used for training. Cross-validation is repeated K-times. Each sub-sample is validate once and the average result of K times is used as the final result. The most common is the 10-fold cross-validation.

To evaluate the performance of the experiment there is the accuracy, which is the proportion of samples that are correctly classified to the total samples.

4.3 Issues of physiological signals

A lot of efforts have been made in revealing the relationships between explicit physiological signals and implicit psychological feelings. How-

ever, there are still several challenges in emotion recognition based on physiological signals:

- Need a very well designed setup in order to obtain high quality physiological data. The standard setup is the standard lab setting with subjects with earphones sit motionless in front of a screen where the emotion stimuli materials are played. This system is fixed, data are noiseless and stable, but the issue of obtaining genuine emotions which is dependent heavily on the emotion stimulated materials is hard to deal with it.

- The stimulus materials are artificially selected, labels of the materials are manually set, while for the same thing human emotions vary from each other, this may lead to a large deviation in the rating of the material. In [48], the author showed that different introduction ways led to different physiological responses.

- There is no still clear evidence that what feature combination of what physiological signal combinations are the most significantly relevant to emotion changes.

- For most studies, the number of subjects is small. Due to limited samples, the performance of the classifier with subjects who have not been trained would be poor. The clear efficient method is to include more subjects from different ages and backgrounds.

- Emotion perception and experience lead to strong differences.

- The reliability of facial expressions cannot be guaranteed sometimes.

4.4 Related works

In this section will be presented some works done during the last years on all physiological signals that tries to understand the relationship of physiology and human emotions.

4.4.1 ECG and GSR signal emotion recognition

In the work [42] Electrocardiogram and Galvanic Skin Response of 11 healthy students were collecred while subjects were listening to emotional music clips. They extracted Matching Pursuit (MP) coefficients frin ECG and GSR signals.

Than were calculated some statistical indices from the MP coefficients and three dimensionality reduction method were applied, PCA, LDA and Kernel PCA. These features were fed into the Probabilistic Neural Network (PNN) in subject-dependent and subject-independent modes.

The PNN is feed with a feature vector in the input layer, is determined a distance between the input and the weight vector. Is calculated the summation of these contributions for each input class to yield the probability. Finally is selected the maximum of these probabilities by a competitive layer and assigned 1 for that class and 0 elsewhere.

They achieved, using PCA, the highest recognition rate of 100% for $\sigma = 0.01$.

4.4.2 ECG sensors for human emotion recognition

The research of [36], they suggest an ensamble learning approach for developing a machine learning model that can recognize four major human emotions (anger, sadness, joy and pleasure) incorporating ECG signals.

As feature extraction method, the analysis combines four ECG signals techniques. Several ML methods have been applied to the model, the most accurate (with an accuracy of 70% is achieved by using an Extra Tree Classifier (a variant of the Random Forest that introduce more variation in the ensemble).

4.4.3 Automatic ECG emotion recognition

In [37] is presented an automatic ECG-based emotion recognition algorithm. They recorded ECG signal from subjects and extracted some features from the signal from the time and frequency domain. Than performed an algoritm of feature selection, a sequential forward floating selection-kernel-based class separability-based.
Valence and arousal and four types of of emotions are recognized using Least Square-Support Vector Machine recognizer. They gained a classification rate for positive/negative valence, high/low arousal, and four types of emotion classification tasks are 82.78%, 72.91%, and 61.52%, respectively.

4.4.4 Classification of music emotions with forehead biosignals and ECG

In the work of [38] to recognize music-induced emotions is used a fusion of three-channel forehead biosignals (left and right temporal channel and frontalis) and ECG. They employed two parallel SVM as arousal and valence classifiers.

The inputs of the classifiers were obtained by applying a fuzzy-rough model feature evaluation criterion and sequential forward floating selection algorithm.
The average classification accuracy was of 88.78% (valence classification accuracy of 94.91% and arousal classification accuracy of 93.63%).

4.4.5 Emotion classification with forehead biosignals

In [39] investigates the feasibility of usign 3-channel forehead bisignals. Classification in vale-arousal space is performed by employing two parallel cascade-forward Neural Networks.
The inputs of the classifiers were obtained by applying a fuzzy rough model feature evaluation criterion and sequential forward floating selection algorithm. An averaged classification accuracy of 87.05% was achieved, corresponding to average valence classification accuracy of 93.66 % and average arousal classification accuracy of 93.29 %.

4.4.6 Physiological changes in music listening

The paper [41] investigates the potential of physiological signals as reliable channels for emotion recognition. Were used four-channel biosensors to measure EMG, ECG, SC and RSP changes as can be seen in figure 4.3. They extracted some features in various analysis domain, the time/frequency, entropy, geometric analysis, subband spectra.

Classification of four musical emotion (one for each quadrant of the valence-arousal diagram) is performed by using an extended linear discriminant analysis (pLDA). They also provided a novel scheme of emotion-specific multilevel dichotomous classification, gaining an accuracy of 95% for subject-dependent and 70% for subject-independent classification.

4.4.7 NN based emotion estimation

In order to build a human-computer interface that is sensitive to a user's expressed emotion, in [43] propose a neural network based emotion estimation algorithm using HR and GSR. In this study, a video clip method was used to elicit basic emotions from subjects while ECG and GSR signals were measured. These signals reflect the influence of emotion on the autonomic nervous system. The extracted features that are emotion-specific characteristics from those signals are applied to an artificial neural network in order to recognize emotions from new signal collections. Results show that the proposed method is able to accurately distinguish a user's emotion.
They gain a total accuracy of 80.2%.

4.4.8 Recognize emotions by affective sound through HRV

The research in [45] reports on how emotional states elicited by affective sounds can be effectively recognized by means of estimates of ANS dynamics.

The ANS dynamics is estimated through standard and nonlinear analysis of HRV exclusively, which is derived from the ECG. A group of 27

people were administered with ECG recordings, then HRV features showing significant changes between valence and arousal dimensions were used as input of an automatic classification system.

The best accuracy was achieved for a quadratic discriminant classifier, to 84.72% on the valence dimension and 84.26% on the arousal one.

4.4.9 Emotion recognition from ECG

In [40] carried out the work of affective ECG signal acquisition from 391 subjects through stimulation of film clips. They recognized emotions divided into Joy and Sadness.
Than, is implemented features extraction and feature selection algorithms based on the DWT and a Fisher-KNN to classify the test data.

4.4.10 Relationship between music emotion and physiological signals

In [5] the study explore the possibility of using physiological signals to detect users emotion response to music, considering individual characteristics (as personality, music preferences, etc.).

A user experiment was conducted with 23 participants, during music listening, a series of physiological signals like HR and SC were recorded using a wearable wristband.
Here, arousal and mood values rated by participants were grouped into three main categories (i.e. positive, negative, neutral), for mood ratings, they combined the mood categories into positive, negative and neutral moods.
After some data preprocessing, were extracted the features in table 4.3:

Category	Features
Descriptive statistics of raw signal	Mean, Standard deviation, median, range
Time series features	Means of the abs of the $1^{st}/2^{nd}$ differences of the raw/normalized signals
Physiological signal specific features	SC response, HR variability

Table 4.3: Features extracted from physiological signals in [5]

A machine learning approach was applied to measure the extent to which physiological signals could be used to recognize users' emotion responses to music listening, in positive and negative categories of arousal and mood. Specifically, they trained and compared the performance of several classification models, namely decision tree, k-Nearest Neighbor, naïve Bayes and SVM.

4.4.11 DL model for human emotion recognition with EDA

The work in [49] had the main objective of ensure that elderly and/or disabled people perform/live well in their immediate environments; this can be monitored by among others the recognition of emotions based on non-highly intrusive sensors such as EDA sensors.
However, designing a learning system or building a machine-learning model to recognize human emotions while training the system on a specific group of persons and testing the system on a totally a new group of persons is still a serious challenge in the field, as it is possible that the second testing group of persons may have different emotion patterns.

They contributed to the field of human emotion recognition by proposing a CNN architecture which promise robustness for both subject-dependent and subject independent human emotion recognition.
Authors converted EDA signals into matrices whereby the goal is to make the application of CNN model possible.
They tested the CNN on two datasets, MAHNOB and DEAP, which are four-classes labeled and they increased the accuracy up to 78% for MAHNOB and 82% for DEAP in subject-independent classification, while up to 81% for MAHNOB and 85% for DEAP in subject-dependent classification.

For the sake of clarity, subject-independent emotion recognition is a challenging field due to:

- Physiological expressions of emotion depend on age, gender, culture and other social factors.
- Depends on the environment in which a subject lives.
- The lab-setting independent nature of emotion recognition is related to the fact that the classifier can/will be trained locally once using sensors of a given lab-setting and after that tested considering different datasets that are collected based on different lab settings.

4.4.12 VA recognition of affective sounds based on EDA

In [50] tried to automatically classify the emotional state of healthy subjects. They proposed the use of convex optimization based on EDA framework and clustering algorithms to automatically discern arousal and valence levels induced by affective stimuli.

EDA recordings were gathered from 25 healthy volunteers, using only one EDA sensor to be placed on fingers.
In model-based approaches, models describe and estimate the underlying psychological process that generates the observed data (EDA measurements). The model based analysis of EDA has fundamental advantages,

such as a propensity to reduce the effects of measurement noise and the essential ability to improve the temporal resolution of inference in rapid event-related paradigms.

EDA data, in this experiment was analyzed with the *cvxEDA* algorithm, presented in [51], which proposed a representation of the SCRs parts of EDA as the ouptut of a linear time-invariant system to a sparse non-negative driver signal.
The model of the *cvxEDA* assumes that the observed SC (y) is the sum of the phasic activity (r), a slow tonic component (t), and an additive independent and identically distributed zero-average Gaussian noise term (ε):

$$y = r + t + \varepsilon \tag{4.1}$$

Physiologically-plausible characteristics (temporal scale and smoothness) of the tonic input signal can be achieved by means of a cubic spline with equally-spaced knots every 10 s, an offset and a linear trend term:

$$t = Bl + Cd \tag{4.2}$$

where B is a tall matrix whose columns are cubic B-spline basis functions, l is the vector of spline coefficients, C is a $Nx2$ matrix with $C_{i,1} = 1$, and $C_{i,2} = \frac{i}{N}$, d is a $2x1$vector with the offset and slope coefficients for the linear trend. Phasic component is the result of a convolution between the sudomotor nerve activity (SMNA) p and an impulse response $h(t)$ shaped as a biexponential Bateman function:

$$h(t) = (e^{-t/\tau_1} - e^{-t/\tau_2})u(t) \tag{4.3}$$

where τ_1 and τ_2 are the slow and the fast time constants of the phasic curve shape and $u(t)$ is the unitary step function.
Referring to [50], the final model can be written as:

$$y = Mq + Bl + Cd + \varepsilon \tag{4.4}$$

Given the EDA model 4.4, *cvxEDA* formulated the problem as a minimization problem as:

$$minimize\ \frac{1}{2}||Mq + Bl + Cd - y||^2 + \alpha\delta||Aq||_1 + \frac{\gamma}{2}||l||_2^2 \tag{4.5}$$

$$subject\ to\ Aq >= 0$$

This problem can be solved using one of the many sparse-QP solvers in order to find the optimal $[q, l, d]$, than find tonic component t from 4.2.
They extracted several features form EDA both from phasic and tonic components output of the *cvxEDA*. For each feature, two levels of valence (positive and negative) and three levels of arousal (low, medium and high) were compared.
The supervised classification was implemented using a kNN classifier.

Results, thanks to *cvxEDA* showed a recognition accuracy of 80 % on the arousal dimension and 84 % in valence classification.

4.5 Conclusions

In this chapter was shown a review of the state of the art about human emotion recognition based on physiological data. A schematic block of the general algorithm can be seen in figure 4.4.
Summing up, the majority of the works deal with a small number of subjects (25/30 maximum) and they evaluated their accuracy based on valence-arousal space grouped in four main labels.

5 Implementation and Results

In this chapter is presented the overall description of the dataset used in the experiment and the related work already done.

5.1 PMEmo dataset

This thesis is based on the paper *The PMEmo dataset for Music Emotion Recognition* [52]. K. Zhang, H. Zhang and S.Li created a novel dataset called *PMEmo* (popular music with emotional annotations) containing emotions of 794 songs as well as EDA signals.
A musical experiment was well-designed for collecting the affective annotated music corpus oh high quality, which recruited 457 subjects. The dataset is publically available to the research comunity at this link[1].
It is intended for benchmarking in MIR and MER, it involves precomputed audio features sets and manually selected chorus excerpts (in .mp3) of songs, to facilitate the development of chorus-related research.

5.1.1 Dataset structure

The dataset contain 794 music clips annotated by 457 subjects, which are subjects from different countries and majors, in order to eliminate the effects of cultural and educational background [53].
Chorus parts are manually selected from students majoring in music.

Meanwhile, the EDA of subjects when listening to these music are also recorded, making it possible to analyses emotion states in multiple

[1] https://drive.google.com/drive/folders/1qDk6hZDGVlVXgckjLq9LvXLZ9EgK9gw0

modes. All annotations are stored in CSV files delimited by comma. The dataset is composed of:

- annotations: valence and arousal values for each song. There are:
  - *static_annotations*: valence and arousal standard deviation values for each song, one value for each song.
  - *static_annotations_std*: valence and arousal mean values for each song, one value for each song.
  - *dynamic_annotations*: valence and arousal standard deviation values for each song, acquired at a sampling rate of $2Hz$.
  - *dynamic_annotations_std*: valence and arousal mean values for each song, acquired at a sampling rate of $2Hz$.
- chorus: all chorus excerpts of 794 songs manually selected.
- comments: songs comments taken from NetEase[2] and SoundCloud[3].
- EDA: EDA data for each song, each one extracted by at least 10 subjects, with a sampling rate of $50Hz$.
- features: all features extracted by the authors of [52]:
  - *EDA_features*: features extracted from the EDA signals:
    - *EDA_features_static*: EDA static features for each song for each subject.
    - *EDA_features_dynamic*: EDA dynamic features for each song for each subject with a sampling rate of $50Hz$.
    - *static_features*: audio static features for each song for each subject.
    - *dynamic_features*: audio dynamic features for each song for each subject with a sampling rate of $50Hz$.
- *lrc_dataset*: lyrics text of all music excerpts.
- lyrics: lyrics text of all music excerpts divided by each timestamp.
- *metadata*: metadata of the songs, containing *music_ID*, title, artist, album, duration, *chorus_start_time* and *chorus_end_time*.

[2]https://music.163.com
[3]https://soundcloud.com

Since the early years of MER, there have been numerous efforts to build datasets with emotional annotations, to facilitate the development and evaluation of music emotion recognition. Table 5.1 from [52] summarize some works on that.

Name	Stimulus	Data	Audio
Emotify[4]	400 excerpts	induced emotion	yes
Moodswing[5]	240 excerpts (30s)	valence and arousal	no
Amg1608[6]	1608 excerpts (30s)	valence and arousal	no
emoMusic[7]	744 excerpts (45s)	valence and arousal	yes
DEAM[8]	1802 excerpts	valence and arousal	yes
SoundTracks[9]	360 + 110 excerpts	valence, energy, tension, mood	yes
GMD[10]	1400 songs	genre, valence and arousal	yes
DEAPDataset[11]	120 music excerpts	valence, arousal, dominance and physiological data	no
PMEmo	794 music chorus	valence, arousal and physiological data	yes

Table 5.1: Some existing music datasets with emotion annotations

5.1.2 Song acquisition and subject selection

They collected 1000 songs from the "*Billboard Hot 100*", the "*iTunes top 100 songs*" and the "*UK top 40 single charts*". They late discovered a set of duplicates and filtered reduplicative music obtaining a full song set of 794 pop songs.
Each datasets in MER utilize music segments, here each clip is manually selected as one of the chorus parts of each song, which is implemented by university students in music major. The clips are of various length, exactly the duration of the chorus parts.

A total of 457 subjects, 236 females and 221 males were recruited to participate. Among them, 366 are Chinese university students who are in non-music majors while 44 are majoring in music recruited to ensure high quality labeling. To weaken the impact of cultural background, 47

[4]http://www.projects.science.uu.nl/memotion/emotifydata/
[5]http://music.ece.drexel.edu/research/emotion/moodswingsturk
[6]https://amg1608.blogspot.ca/
[7]http://cvml.unige.ch/databases/emoMusic/
[8]http://cvml.unige.ch/databases/DEAM/
[9]https://www.jyu.fi/hytk/fi/laitokset/mutku/en/research/projects2/pastprojects/coe/materials/emotion/soundtracks/Index
[10]https://hilab.di.ionio.gr/en/music-information-research/
[11]http://www.eecs.qmul.ac.uk/mmv/datasets/deap/readme.html

English speakers are invited to annotate the datasets.
Each song received a total of at least 10 emotion annotations including one by music-majoring and one by English speaker.

5.1.3 Experiment design

To monitor and obtain EDA continuously they used MP150 Biopac system[12] at a sampling rate of $50Hz$ and export signals from *AcqKnowledge* software.
To annotate songs was developed a desktop application shown in figure 5.1. The annotation was done with the sliding area collecting dynamics annotations , from 1 to 9, at a sampling rate of $2Hz$. Annotators should make a statistic annotation for the whole music excerpts on nine-point scale after the dynamic labeling. Furthermore, annotators were asked to listen to the same music twice to annotate on valence and arousal separately.

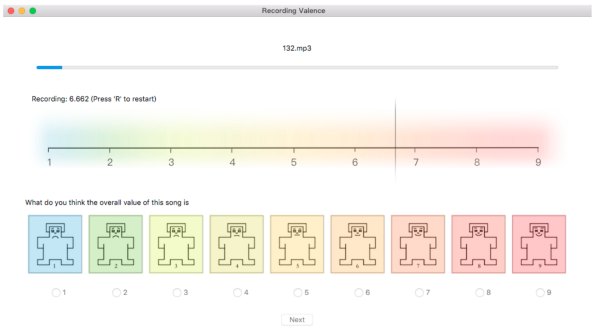

Figure 5.1: Annotation interface for PMEmo

In figure 5.2 is shown the flow diagram of the experiment, where each subject spent 50 minutes on average.

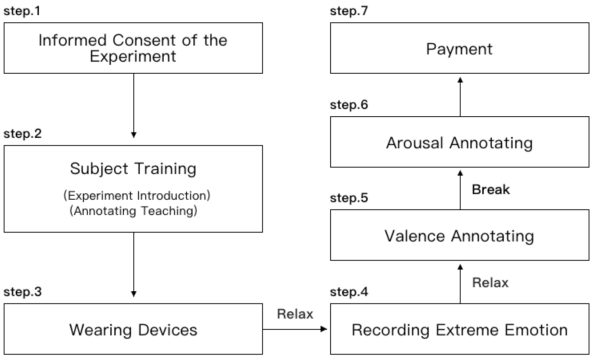

Figure 5.2: Experimental procedure for PMEmo

Each subject listened to 20 excerpts and one of those was duplicated to guarantee the high quality data as what was done in [54]. The annotations from this subjects were accepted only if the bias between duplicate

[12]https://www.biopac.com/product/eda-finger-transducer-bsl/

clips were within 0.25 in the valence-arousal space (they did not inform subjects about the duplicated excerpt).
In total 457 subjects have participated but 401 were considered valid annotations (87.7%). Each music clip was annotated by at least 10 subjects including English speakers and semi-experts from music academy.

5.1.4 Data reliability

As Initial Orientation Time (IOT) concept remarks, annotators need some preliminary time before they can give meaningful and reliable annotations. Schubert in [55] found that median IOT for valence was $8s$ while for arousal $12s$. Other researchers showed that annotations began to converge after $10s$. The PMEmo authors decided to discard first $15s$ for the dynamic annotations from the data.

To evaluate annotation consistency they used the Chronbach's α, it represent the degree to which a set of items measures a single unidimensional latent construct. In [52] computed the Chronbach's α on the sequence of annotations for each song.
They processed annotations by:

$$aj, i = aj, i + (\bar{A}_j - \bar{A}) \tag{5.1}$$

where:

- aj, i is the label annotated by subject j at time i
- $\bar{A}$ is the mean of all the labels for this song by all subjects
- $\bar{A}_j$ is the mean of dynamic labels by subject j

6
Dataset Improvements

In this chapter we present...
Then...
Finally ...

6.1 Some different sections

6.2 Conclusive Remarks

7 Conclusions and Future Works

This work of thesis proposes a methodology for...
The devised methodology is based on...
The main advantages are...
As far as the experiments are concerned...
The proposed approach has shown promising results both in simulation and in the experiments.

7.1 Future Works

Generalization We would like to generalize...

Challenging scenarios Another possible improvement is related to the extension of the proposed approach to...

Different approaches Finally we are moving towards a deeper analysis of... a a a a a a a a a a a a a a

Appendices

Bibliography

[1] Y. Feng, Y. Zhuang, and Y. Pan, "Popular music retrieval by detecting mood," in *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, (Hangzhou, China), pp. 375–376, 2003.

[2] E. A. Kensinger, "Remembering emotional experiences: The contribution of valence and arousal," *Reviews in the Neurosciences*, vol. 15, no. 4, pp. 241–252, 2004.

[3] R. Panda, R. M. Malheiro, and R. P. Paiva, "Novel audio features for music emotion recognition," *IEEE Transactions on Affective Computing*, 2018.

[4] J. Shukla, M. Barreda-Angeles, J. Oliver, G. Nandi, and D. Puig, "Feature extraction and selection for emotion recognition from electrodermal activity," *IEEE Transactions on Affective Computing*, 2019.

[5] X. Hu, F. Li, and T.-D. J. Ng, "On the relationships between music-induced emotion and physiological signals," in *ISMIR*, pp. 362–369, 2018.

[6] Y.-H. Yang and H. H. Chen, *Music emotion recognition*. USA: CRC Press, Inc., 1st ed., 2011.

[7] J. H. Lee and J. S. Downie, "Survey of music information needs, uses, and seeking behaviours: preliminary findings," in *ISMIR*, vol. 2004, p. 5th, Citeseer, 2004.

[8] P. N. Juslin and P. Laukka, "Expression, perception, and induction of musical emotions: A review and a questionnaire study of everyday listening," *Journal of new music research*, vol. 33, no. 3, pp. 217–238, 2004.

[9] B. Van De Laar, "Emotion detection in music, a survey," in *Twente Student Conference on IT*, vol. 1, p. 700, 2006.

[10] K. Hevner, "Expression in music: a discussion of experimental studies and theories," *Psychological review*, vol. 42, no. 2, p. 186, 1935.

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

[12] X. Yang, Y. Dong, and J. Li, "Review of data features-based music emotion recognition methods," *Multimedia Systems*, vol. 24, no. 4, pp. 365–389, 2018.

[13] K. F. MacDorman, Stuart Ough Chin-Chang Ho, "Automatic emotion prediction of song excerpts: Index construction, algorithm design, and empirical comparison," *Journal of New Music Research*, vol. 36, no. 4, pp. 281–299, 2007.

[14] A. Gabrielsson and E. Lindström, "The influence of musical structure on emotional expression," 2001.

[15] J. L. Zhang, X. L. Huang, L. F. Yang, Y. Xu, and S. T. Sun, "Feature selection and feature learning in arousal dimension of music emotion by using shrinkage methods," *Multimedia Systems*, vol. 23, no. 2, pp. 251–264, 2017.

[16] M. Robnik-Šikonja and I. Kononenko, "Theoretical and empirical analysis of relieff and rrelieff," *Machine learning*, vol. 53, no. 1-2, pp. 23–69, 2003.

[17] A. Géron, *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. O'Reilly Media, 2019.

[18] W. Boucsein, *Electrodermal activity*. Springer Science & Business Media, 2012.

[19] W. Trost, T. Ethofer, M. Zentner, and P. Vuilleumier, "Mapping aesthetic musical emotions in the brain," *Cerebral Cortex*, vol. 22, no. 12, pp. 2769–2783, 2012.

[20] L. C. Johnson and A. Lubin, "Spontaneous electrodermal activity during waking and sleeping," *Psychophysiology*, vol. 3, no. 1, pp. 8–17, 1966.

[21] D. C. Fowles, "Electrodermal activity and antisocial behavior: Empirical findings and theoretical issues," in *Progress in electrodermal research*, pp. 223–237, Springer, 1993.

[22] A. Bizzego, A. Battisti, G. Gabrieli, G. Esposito, and C. Furlanello, "pyphysio: A physiological signal processing library for data science approaches in physiology," *SoftwareX*, vol. 10, p. 100287, 2019.

[23] J. Hernandez, I. Riobo, A. Rozga, G. D. Abowd, and R. W. Picard, "Using electrodermal activity to recognize ease of engagement in children during social interactions," in *Proceedings of the 2014*

ACM International Joint Conference on Pervasive and Ubiquitous Computing, pp. 307–317, 2014.

[24] M. Benedek and C. Kaernbach, "A continuous measure of phasic electrodermal activity," *Journal of neuroscience methods*, vol. 190, no. 1, pp. 80–91, 2010.

[25] S. Taylor, N. Jaques, W. Chen, S. Fedor, A. Sano, and R. Picard, "Automatic identification of artifacts in electrodermal activity data," in *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 1934–1937, IEEE, 2015.

[26] E. B. Hedman, *In-situ measurement of electrodermal activity during occupational therapy.* PhD thesis, Massachusetts Institute of Technology, 2010.

[27] A. Drachen, L. E. Nacke, G. Yannakakis, and A. L. Pedersen, "Correlation between heart rate, electrodermal activity and player experience in first-person shooter games," in *Proceedings of the 5th ACM SIGGRAPH Symposium on Video Games*, pp. 49–54, 2010.

[28] J. A. Healey and R. W. Picard, "Detecting stress during real-world driving tasks using physiological sensors," *IEEE Transactions on intelligent transportation systems*, vol. 6, no. 2, pp. 156–166, 2005.

[29] P. Ghaderyan and A. Abbasi, "An efficient automatic workload estimation method based on electrodermal activity using pattern classifier combinations," *International Journal of Psychophysiology*, vol. 110, pp. 91–101, 2016.

[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] S. De Nadai, M. D'Incà, F. Parodi, M. Benza, A. Trotta, E. Zero, L. Zero, and R. Sacile, "Enhancing safety of transport by road by online monitoring of driver emotions," in *2016 11th System of Systems Engineering Conference (SoSE)*, pp. 1–4, Ieee, 2016.

[32] R. Guo, S. Li, L. He, W. Gao, H. Qi, and G. Owens, "Pervasive and unobtrusive emotion sensing for human mental health," in *2013 7th International Conference on Pervasive Computing Technologies for Healthcare and Workshops*, pp. 436–439, IEEE, 2013.

[33] B. Verschuere, G. Crombez, E. Koster, and K. Uzieblo, "Psychopathy and physiological detection of concealed information: a review," *Psychologica Belgica*, vol. 46, no. 1-2, 2006.

[34] Y.-D. Zhang, Z.-J. Yang, H.-M. Lu, X.-X. Zhou, P. Phillips, Q.-M. Liu, and S.-H. Wang, "Facial emotion recognition based on biorthogonal wavelet entropy, fuzzy support vector machine, and stratified cross validation," *IEEE Access*, vol. 4, pp. 8375–8385, 2016.

[35] Q. Mao, M. Dong, Z. Huang, and Y. Zhan, "Learning salient features for speech emotion recognition using convolutional neural networks," *IEEE transactions on multimedia*, vol. 16, no. 8, pp. 2203–2213, 2014.

[36] T. Dissanayake, Y. Rajapaksha, R. Ragel, and I. Nawinne, "An ensemble learning approach for electrocardiogram sensor based human emotion recognition," *Sensors*, vol. 19, no. 20, p. 4495, 2019.

[37] Y.-L. Hsu, J.-S. Wang, W.-C. Chiang, and C.-H. Hung, "Automatic ecg-based emotion recognition in music listening," *IEEE Transactions on Affective Computing*, 2017.

[38] M. Naji, M. Firoozabadi, and P. Azadfallah, "Classification of music-induced emotions based on information fusion of forehead biosignals and electrocardiogram," *Cognitive Computation*, vol. 6, no. 2, pp. 241–252, 2014.

[39] M. Naji, M. Firoozabadi, and P. Azadfallah, "Emotion classification during music listening from forehead biosignals," *Signal, Image and Video Processing*, vol. 9, no. 6, pp. 1365–1375, 2015.

[40] J. Cai, G. Liu, and M. Hao, "The research on emotion recognition from ecg signal," in *2009 International Conference on Information Technology and Computer Science*, vol. 1, pp. 497–500, IEEE, 2009.

[41] J. Kim and E. André, "Emotion recognition based on physiological changes in music listening," *IEEE transactions on pattern analysis and machine intelligence*, vol. 30, no. 12, pp. 2067–2083, 2008.

[42] A. Goshvarpour, A. Abbasi, and A. Goshvarpour, "An accurate emotion recognition system using ecg and gsr signals and matching pursuit method," *biomedical journal*, vol. 40, no. 6, pp. 355–368, 2017.

[43] S. K. Yoo, C. K. Lee, Y. J. Park, N. H. Kim, B. C. Lee, and K. S. Jeong, "Neural network based emotion estimation using heart rate variability and skin resistance," in *International conference on natural computation*, pp. 818–824, Springer, 2005.

[44] O. Sourina, Y. Liu, and M. K. Nguyen, "Real-time eeg-based emotion recognition for music therapy," *Journal on Multimodal User Interfaces*, vol. 5, no. 1-2, pp. 27–35, 2012.

[45] M. Nardelli, G. Valenza, A. Greco, A. Lanata, and E. P. Scilingo, "Recognizing emotions induced by affective sounds through heart

rate variability," *IEEE Transactions on Affective Computing*, vol. 6, no. 4, pp. 385–394, 2015.

[46] D. M. Sloan, "Emotion regulation in action: Emotional reactivity in experiential avoidance," *Behaviour Research and Therapy*, vol. 42, no. 11, pp. 1257–1270, 2004.

[47] S. Jerritta, M. Murugappan, R. Nagarajan, and K. Wan, "Physiological signals based human emotion recognition: a review," in *2011 IEEE 7th International Colloquium on Signal Processing and its Applications*, pp. 410–415, IEEE, 2011.

[48] C. Mühl, A. Brouwer, N. van Wouwe, E. van den Broek, F. Nijboer, and D. K. Heylen, *Modality-specific affective responses and their implications for affective BCI*. Graz, Austria: Verlag der Technischen Universität, 2011.

[49] F. Al Machot, A. Elmachot, M. Ali, E. Al Machot, and K. Kyamakya, "A deep-learning model for subject-independent human emotion recognition using electrodermal activity sensors," *Sensors*, vol. 19, no. 7, p. 1659, 2019.

[50] A. Greco, G. Valenza, L. Citi, and E. P. Scilingo, "Arousal and valence recognition of affective sounds based on electrodermal activity," *IEEE Sensors Journal*, vol. 17, no. 3, pp. 716–725, 2016.

[51] A. Greco, G. Valenza, A. Lanata, E. P. Scilingo, and L. Citi, "cvxeda: A convex optimization approach to electrodermal activity processing," *IEEE Transactions on Biomedical Engineering*, vol. 63, no. 4, pp. 797–804, 2015.

[52] K. Zhang, H. Zhang, S. Li, C. Yang, and L. Sun, "The pmemo dataset for music emotion recognition," in *Proceedings of the 2018 acm on international conference on multimedia retrieval*, pp. 135–142, 2018.

[53] X. Hu and Y.-H. Yang, "Cross-dataset and cross-cultural music mood prediction: A case on western and chinese pop songs," *IEEE Transactions on Affective Computing*, vol. 8, no. 2, pp. 228–240, 2017.

[54] Y.-A. Chen, Y.-H. Yang, J.-C. Wang, and H. Chen, "The amg1608 dataset for music emotion recognition," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 693–697, IEEE, 2015.

[55] E. Schubert, "Reliability issues regarding the beginning, middle and end of continuous emotion ratings to music," *Psychology of music*, vol. 41, no. 3, pp. 350–371, 2013.