BABEŞ BOLYAI UNIVERSITY, CLUJ NAPOCA, ROMÂNIA
FACULTY OF MATHEMATICS AND COMPUTER SCIENCE

Automatic self-assessment tool

– MIRPR report –

Mirt Leonard
Computer Science, 234, leonard.mirt@stud.ubbcluj.ro

Stancu Diana-Elena
Computer Science, 236, diana.stancu@stud.ubbcluj.ro

Tofan Raul
Computer Science, 237, cristian.tofan@stud.ubbcluj.ro

Vaida Radu
Computer Science, 237, radu.vaida@stud.ubbcluj.ro

2023-2024

Abstract

This document presents our response (and solution) to the lack of an automated self-assessment tool that could lead to performance improvement in the context of human-to-human interactions such as formal or informal meetings and interviews. In order to solve the issue raised by the NATO HUMINT Centre of Excellence, we developed an emotion recognition tool. Our emotion recognition application employs intelligent methods to analyze speech, video, and physiological signals from a digital watch. Every direction has a distinct pipeline, and by utilising our models, which were influenced by articles recently published and the current state of the art, **we achieved great accuracy with significant growth potential**. Each of these models are integrated into a web application that performs three separate analyses before aligning the findings with a conclusion.

Contents

List of Figures

Chapter 1

Introduction

1.1 What? Why? How?

The scientific problem at the core of our research lies in the nuanced realm of emotion recognition across multiple modalities: speech, video, and pulse from the watch sensor. Understanding and accurately interpreting human emotions have become increasingly vital, spanning applications in healthcare, human-computer interaction and human resources. The challenge is to develop a comprehensive approach that integrates these disparate data sources intelligently. Our basic approach involves leveraging advanced machine learning algorithms to decode emotional cues from audio inputs, facial expressions, and physiological signals. This research significantly contributes to related work by providing a holistic framework for emotion recognition, moving beyond isolated modalities. The objectives of this documentation include refining the integration of diverse data sources, optimizing deep learning models, and demonstrating the practical implications of our approach. The anticipated results aim to showcase heightened accuracy in emotion recognition.

1.2 Paper structure and original contribution(s)

Our emotion recognition application is a fresh take on combining audio, video, and watch sensor inputs. Our unique contribution lies in the way we approach and blend the results. Through experimentation, we've aligned the models in an unique way. We reduced/increased the confidence of a particular emotion resulted from audio/video input based on the confidence derived from physiological emotion analyses (e.g. if the output from audio and video indicates that a person is calm but the output from physiological sensors indicates otherwise, the confidence is recalibrated using penalizations and rewards 4.8).

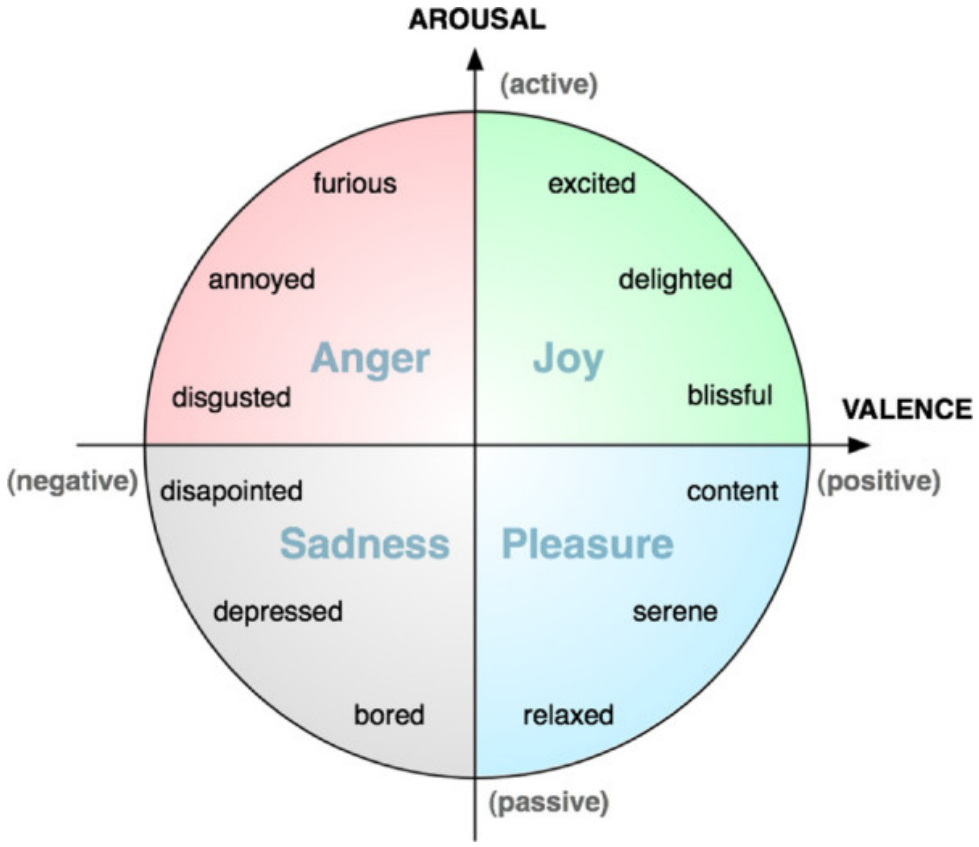


Figure 1.1: Emotions

The document is structured as follows:

Introduction: This section outlines the paper's objectives and structure, highlighting original contributions.

Scientific Problem: This part defines the scientific problem being addressed.

State of the Art/Related Work [3]: This section reviews existing research and relevant work in the field.

Investigated Approach [4]: This segment delves into the approach, breaking it down into sections such as Face Emotion Recognition, Audio Emotion Recognition, and Emotion Recognition based on Physiological Signals. Each subsection includes an introduction, methodology, and model details.

Application (Study Case) [5]: This part describes an application, covering functionalities, design, implementation, testing, numerical validation methodology, data, results, and discussion.

Conclusion and Future Work [6]: The document concludes by summarizing findings and suggesting potential directions for future research.

Chapter 2

Scientific Problem

For both theoretical and practical reasons researchers define emotions according to one or more dimensions. Dimensional models of emotion attempt to conceptualize human emotions by defining where they lie in two or three dimensions. Most dimensional models incorporate valence and arousal or intensity dimensions. The circumplex model of emotion was developed by James Russell.[5] This model suggests that emotions are distributed in a two-dimensional circular space, containing arousal and valence dimensions. Arousal represents the vertical axis and valence represents the horizontal axis, while the center of the circle represents a neutral valence and a medium level of arousal.[6] In this model, emotional states can be represented at any level of valence and arousal, or at a neutral level of one or both of these factors. Circumplex models have been used most commonly to test stimuli of emotion words, emotional facial expressions, and affective states.[7]

Russell and Lisa Feldman Barrett describe their modified circumplex model as representative of core affect, or the most elementary feelings that are not necessarily directed toward anything. Different prototypical emotional episodes, or clear emotions that are evoked or directed by specific objects, can be plotted on the circumplex, according to their levels of arousal and pleasure.

2.1 Problem definition

The problem at hand centers around the complex nature of emotion recognition, a multifaceted task requiring the synthesis of information from speech, video, and watch sensor data. Traditional methods often fall short due to the intricate and dynamic interplay of these modalities, necessitating the application of intelligent algorithms. Conversely, the challenge lies in optimizing these algorithms to handle the diverse input streams efficiently. Formally, the problem involves developing an intelligent system that takes speech, video, and physiological signals as inputs and produces accurate emotional

classifications as outputs. This problem is of paramount importance due to its widespread applicability and successfully addressing this challenge not only advances the field of emotion recognition but also has far-reaching implications for enhancing human-machine interactions across various domains.

Chapter 3

State of the art/Related work

Studies on emotion recognition can be different in many ways, e.g., they may vary in: the dataset used; the emotional model applied; the machine learning approach adopted; the number of classification classes and their distribution; the validation strategy (e.g., user-independent vs. user-dependent); whether the results are provided for train, validation, or test set; and the performance quality measure, thus is difficult to compare them. Awais et al. applied LSTM architecture to classify four emotions (amusement, boredom, relaxation, and fear) based on physiological signals [2]. Their multimodal approach achieved results above 93% of F-measure value. They used CASE [7], a publicly available dataset with a rich spectrum of signals, i.e., ECG, BVP, GSR, RSP, SKT, and EMG. Tizzano et al. used LSTM to recognize happy, sad, and neutral affective states in subjects that were listening to music or watching a short movie [8]. They claimed to obtain 99% accuracy. Dar et al. utilized CNN and LSTM to classify high/low arousal/valence from EEG, ECG, and EDA signals [3]. They claimed 91% and 99% accuracy on DREAMER and AMIGOS datasets, respectively.

An important aspect to consider in the emotion recognition task is model generalization and personalization. On the one hand, we would like to have a predictive model that can be applied irrespectively to the participantsâ demographic and physiological characteristics. This would allow recognizing emotions in people that did not provide any prior data or are interacting with the system for the first time. On the other hand, people are very different in terms of psychological and physiological elements when it comes to expressing and perceiving emotions. Hence, a personal model might perform better than the general one.

Chapter 4

Investigated approach

Through this chapter we will analyse the 3 models that we have built separately (video, audio and physiological data).

4.1 Face emotion recognition

4.1.1 Introduction

Face Emotion Recognition refers to identifying basic emotions (such as angry, disgust, happy, sad, etc) of a human from an image.

4.1.2 Methodology

4.1.2.1 Datasets

- Facial Expression Recognition 2013 (FER2013) (max accuracy: 92.5%)

  Consists of 35888 images of 7 different emotions: anger, disgust, fear, happiness, neutral, sadness, surprise.

  In a Kaggle forum discussion facilitated by the competition organizers, it was mentioned that human accuracy on this dataset falls within the range of 65% to 68%.[4]

- Facial Expressions Training Data

  Based on AffectNet-HQ, which used a state-of the art model to improve the original labels

  The dataset contains 28175 images of 8 emotions: anger, contempt, disgust, fear, happy, neutral, sad, surprise. We used just the images that are labelled with the emotions present in the FER-2013 dataset

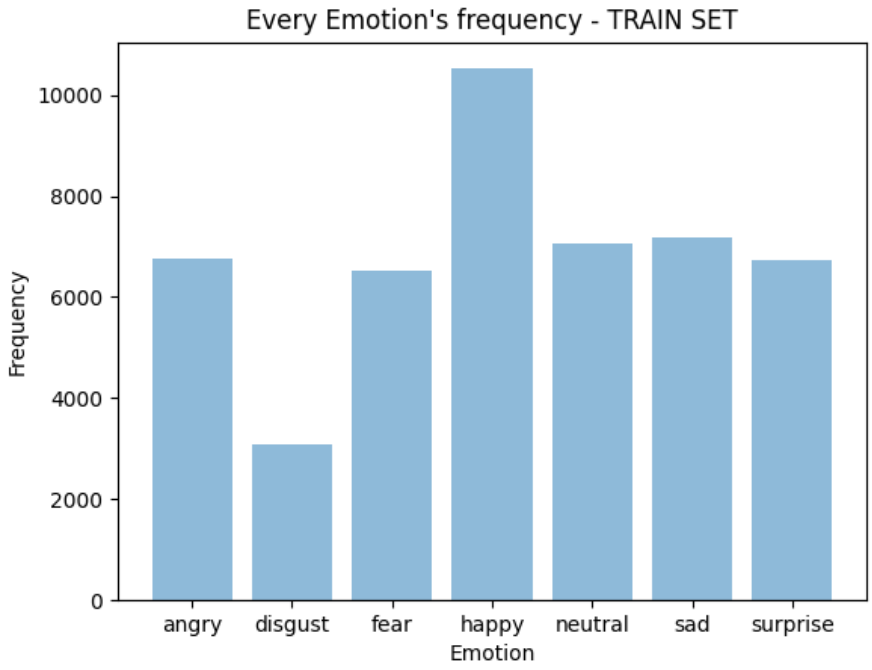


Figure 4.1: Visualise dataset emotions and the number of images for every emotion

4.1.2.2 Data preparation

Our dataset consists of the two merged datasets presented in the "Datasets" subsection. The information is represented as numpy arrays of size 48x48x3. The labels for the arrays are structured in a categorical form, providing an organized way to represent different emotions.

4.1.2.3 Data augmentation

Various augmentation techniques, including rotation, zoom, and horizontal flip, are applied to generate additional data for the model. This helps prevent overfitting during training.

4.1.2.4 MobileNetV1

CNNs have been used very often for identifying objects in images, or classifying images. However, CNNs represent a small disadvantage, they usually require a lot of calculations and they need very large datasets. To solve this problem, MobileNetV1 was created. An algorithm that promises to solve this problem, especially for mobile or embedded devices. Using depthwise separable convolutions, the model manages to greatly reduce the number of parameters that a CNN needs. Inspired by the PAtt-Lite model[6], we used MobileNetV1 as the main feature extractor.

4.1.2.5 Patch Extraction

Our model uses just the first 11 convolution blocks from MobileNetV1 and after them, we added a Patch Extraction block inspired by PAtt-Lite model[6]. The output of the truncated MobileNet is padded with zero to create a new feature map of size 16x16. Our patch extraction block is made actually of three smaller blocks. Specifically, the first two blocks employ depthwise separable convolution followed by batch normalization, along with max pooling. The final block is represented by a pointwise separable convolution.

4.1.2.6 Global Average Pooling

First introduced in [5], Global Average Pooling (GAP) was created to reduce the overfitting problem in CNNs. In our case, Global Average Pooling was used to average the representations provided by the last layer in our Patch Extraction block. We used GAP, not only because it helps to reduce overfitting, but also because it reduces the number of model parameters.

4.1.2.7 Attention Classifier

Dot product attention is a type of self-attention mechanism. It calculates attention weights by taking the dot product of the query vector and the key vector, then dividing the result by the square root of the key vector's dimension. We incorporated an attention classifier between two fully connected layers, enhancing our model's adaptability to the training data.

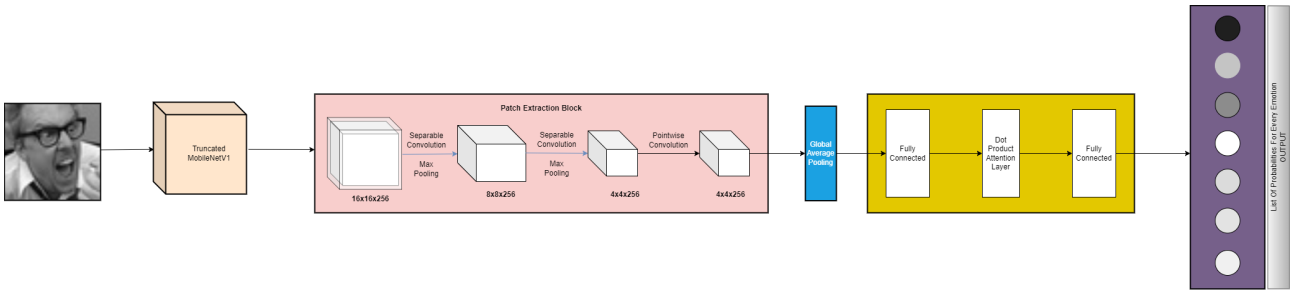


Figure 4.2: Visualize how our proposed model works

4.1.3 Model

Our model got an 80.1% test accuracy, which is quite impressive, taking into account the number of parameters (2.03M) that our model has. We used Adam as an optimizer, with a learning rate of 1e-3. We also, used Reduce Learning Rate On Plateau. It helped us reduce the overfitting problem and also improved our accuracy. In Figure 3, we can see, that even if we had a very few images to train on

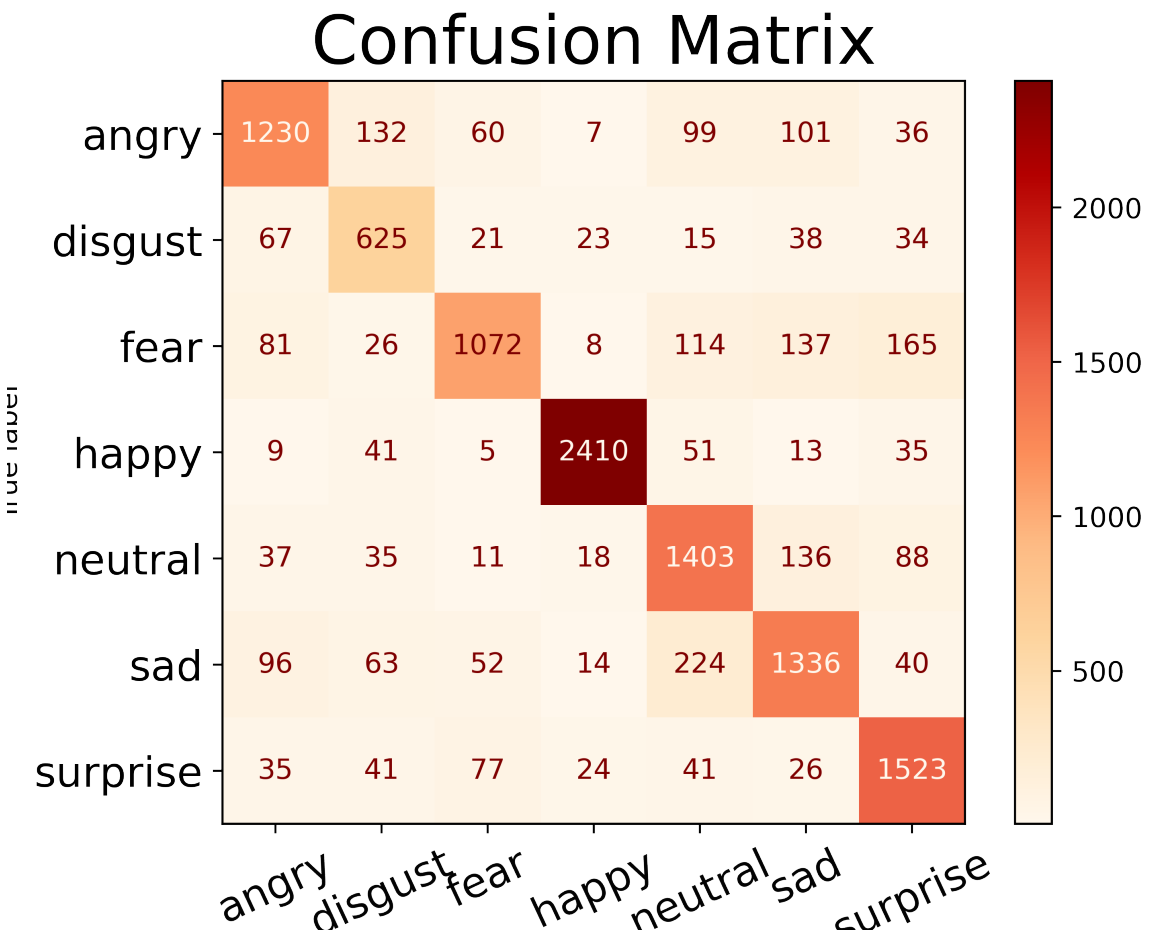


Figure 4.3: Confusion Matrix for our best model

COMPARISON CHART

No of experiment	Optimizer	Learning Rate	Reduce Learning on Plateau decrasing factor	No of epoch of no improvement for reducing learning rate	Epsilon	Clipnorm	Class Weights	Dropout for MobileNetV1	Dropout for Self Attention Layer	Accuracy
1	Adam	1e-3	0.1	3	1e-7(default)	1.0	Yes	0.1	0	69.2%
2	Adam	1e-3	0.1	3	1e-7(default)	1.5	Yes	0.1	0	69.7%
3	Adam	1e-3	0.1	3	1e-2	1.3	Yes	0.1	0	70.6%
4	Adam	1e-3	0.1	3	1e-2	1.35	Yes	0.1	0.2	79.1%
5	Adam	1-e3	0.5	3	1e-2	1.35	Yes	0.1	0.2	79.9%
6	**Adam**	**1e-3**	**0.5**	**3**	**1e-2**	**1.35**	**Yes**	**0.2**	**0.2**	**80.1%**
7	Adam	1e-3	0.5	5	1e-2	1.35	Yes	0.2	0.2	78.59

Figure 4.4: Experiments run on current architecture

disgust class (in comparison with the other classes), we got an impressive accuracy. 625 of 823 (75.9%) images were correctly classified as 'disgust'.

4.1.3.1 Experiments on current architecture

During the experimental phase, we adjusted multiple hyperparameters, like: the learning rate, the clipnorm and the epsilon for the Adam optimizer, the dropout rate in MobileNet and in the attention layer. We will describe the changes that we've done and the accuracy that we've got in the table below.

4.1.3.2 Experiments on older architectures

Older architectures that we used, were trained just on FER-2013. We tried two ideas to solve the unbalanced data: augmenting data every epoch and using class weights. Unfortunately, these solutions are not so good as using two datasets to train our model on. During that time, we tried to find the optimum architecture, but we were also learning. So, we will describe in this section the architectures that we used and the hyperparameters that we adjusted during the training.

4.1.3.3 Simple CNN Architecture

The first architecture that was used to solve the problem was a simple CNN with four convolution block and two fully connected layers. A convolution block was made of a convolution layer, a max pooling layer and a dropout layer. The first convolution had 64 filters, the second had 128 filters, the third one had 256 filters and the last one 512 filters. The maximum accuracy this model got is 58%.

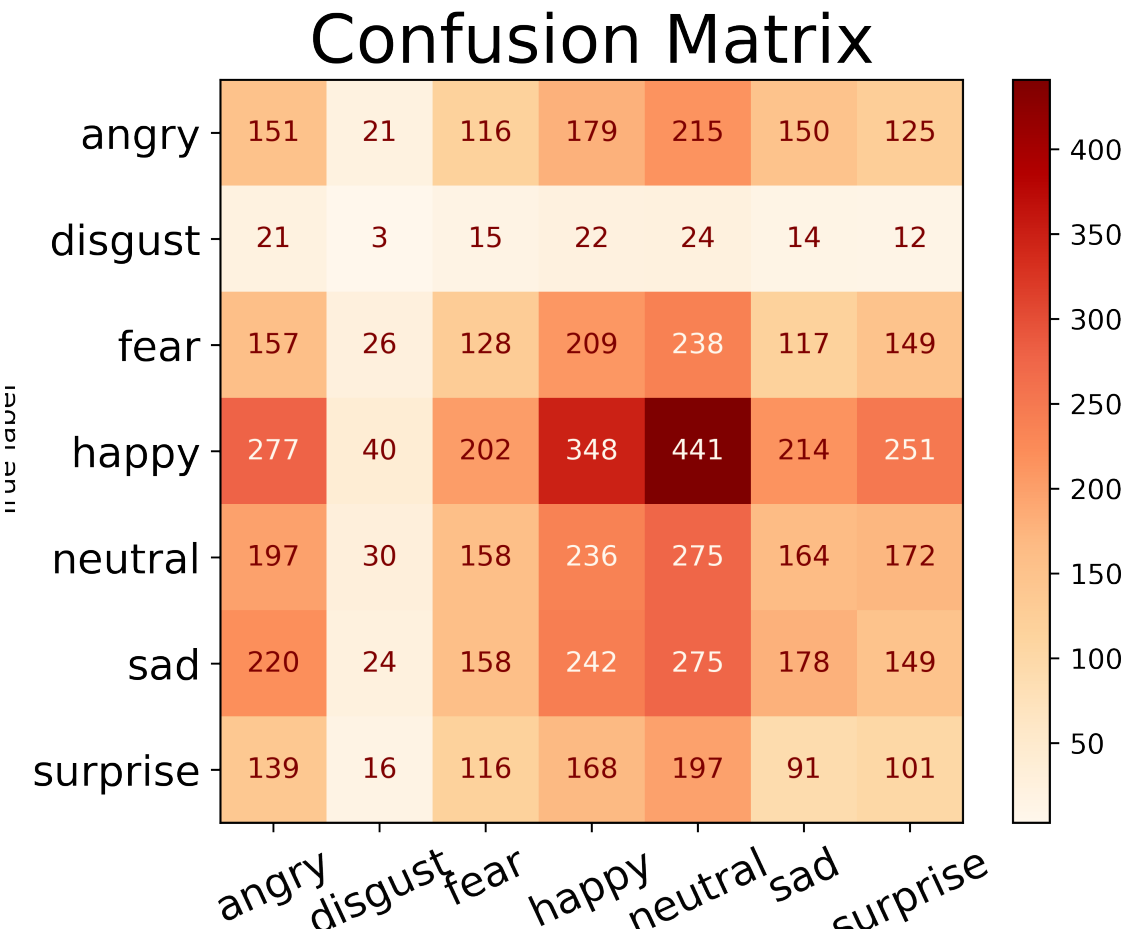


Figure 4.5: Confusion Matrix on our Simple CNN Architecture

An architecture inspired by InceptionV3: Inspired by the InceptionV3 architecture, we tried to create a similar architecture, but with less parameters. It was one of the first more complex models that we trained and the confusion matrix (Fig. 6) was a disaster, but we were proud of what we've done because this model got a 60% accuracy.

An architecture inspired by MobileNetV1: While reading about MobileNetV1, we saw that their architecture uses a lot of depthwise convolutions. So, we created an architecture that uses a

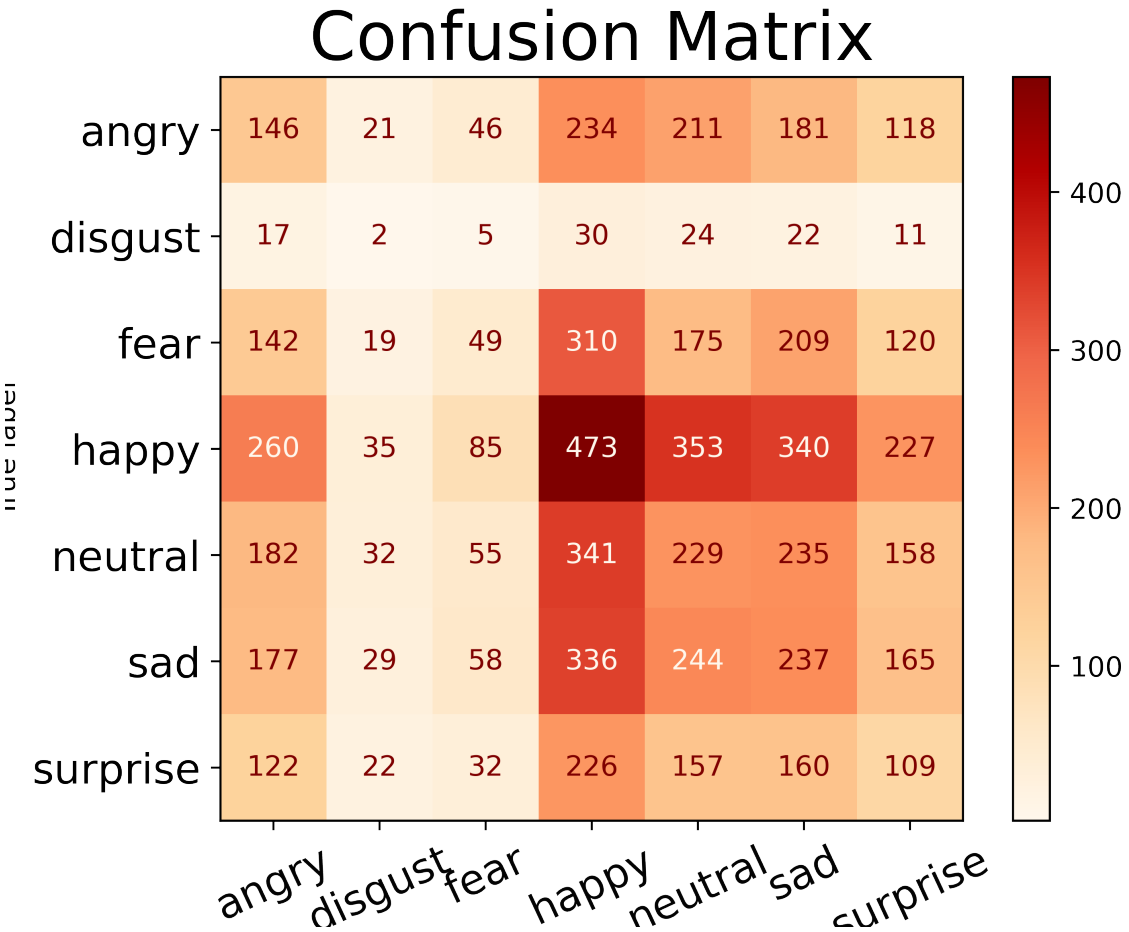


Figure 4.6: Confusion Matrix on InceptionV3 inspired architecture

lot of depthwise convolutions and with 1.64M parameters, we got a 65% accuracy, which is pretty decent taking into account the small number of parameters. The confusion matrix (Fig. 7) has shown noticeable improvement in its visual representation. The most significant improvement was achieved through the utilization of the Image Data Generator (from Tensorflow [1]) because it generates a new set of artificial data every epoch of the training, so it reduces the overfitting. For the first time, we equalized the human accuracy, so it was a very good result.

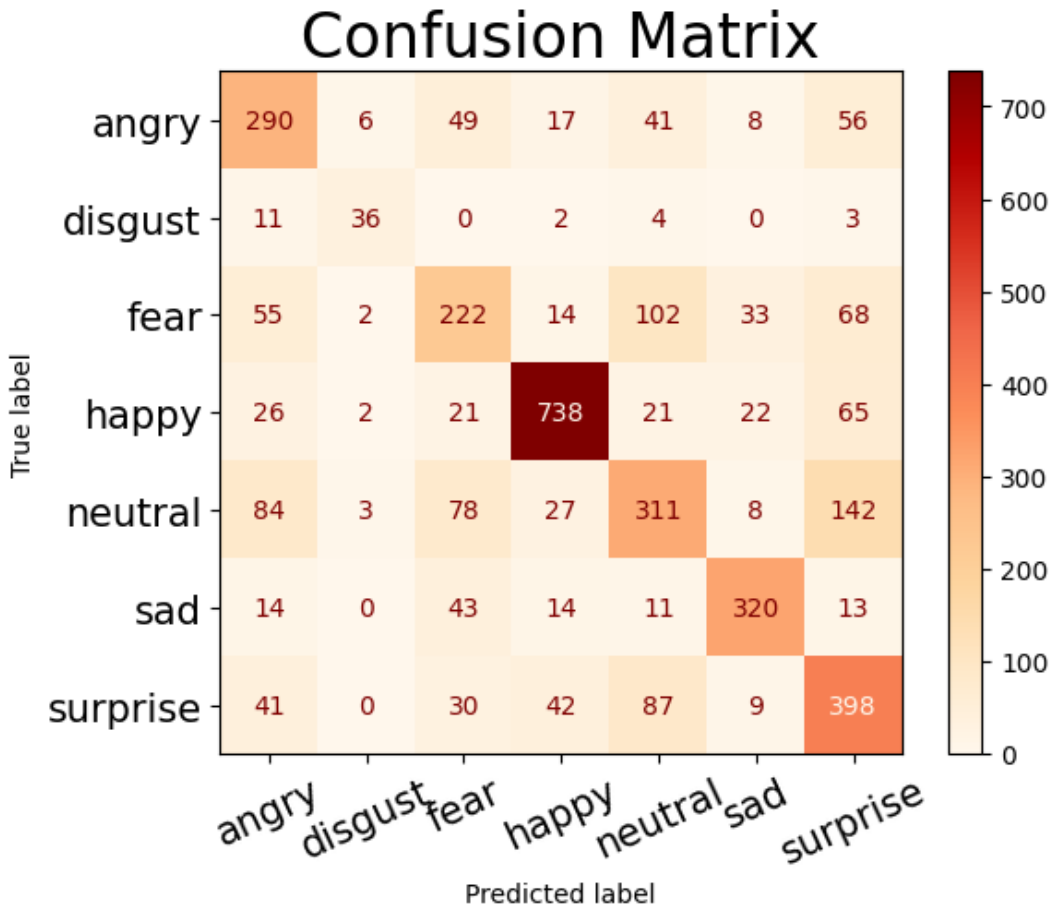


Figure 4.7: Confusion Matrix on MobileNetV1 inspired architecture

4.2 Audio emotion recognition

4.2.1 Introduction

Speech Emotion Recognition, abbreviated as SER, is the act of attempting to recognize human emotion and affective states from speech. This is capitalizing on the fact that voice often reflects underlying emotion through tone and pitch. This is also the phenomenon that animals like dogs and horses employ to be able to understand human emotion.

Audio speech emotion recognition can be approached through two main directions. The first involves analyzing the emotional meaning of words within the speech. For instance, phrases like "I like this beautiful day" convey positivity, while expressions like "I hate this problem. I feel very sad today" denote negativity based on the semantic content of the words. The second direction focuses on recognizing emotions from non-verbal elements present in speech. Non-verbal elements include aspects such as pitch, tone, intonation, speech rate, and pauses. We opted for the second method due to its capacity to capture a more comprehensive range of emotional cues, encompassing subtle nuances in speech delivery that may not be fully conveyed through the semantic analysis of words alone. This approach allows for a more holistic and nuanced understanding of emotional expressions in spoken language.

4.2.2 Methodology

4.2.2.1 Datasets

- Crowd-sourced Emotional Mutimodal Actors Dataset (Crema-D) (max accuracy: 82%)
- Ryerson Audio-Visual Database of Emotional Speech and Song (Ravdess) (max accuracy: 84%)
- Surrey Audio-Visual Expressed Emotion (Savee) (max accuracy: 82%)

4.2.2.2 Data preparation

A dataframe is created for storing all emotions of the data and is used to extract features for our model training.

4.2.2.3 Data augmentation

We use 3 of augmentation techniques noise, stretching(ie. changing speed) and pitching since we can assure our training model is not overfitted.

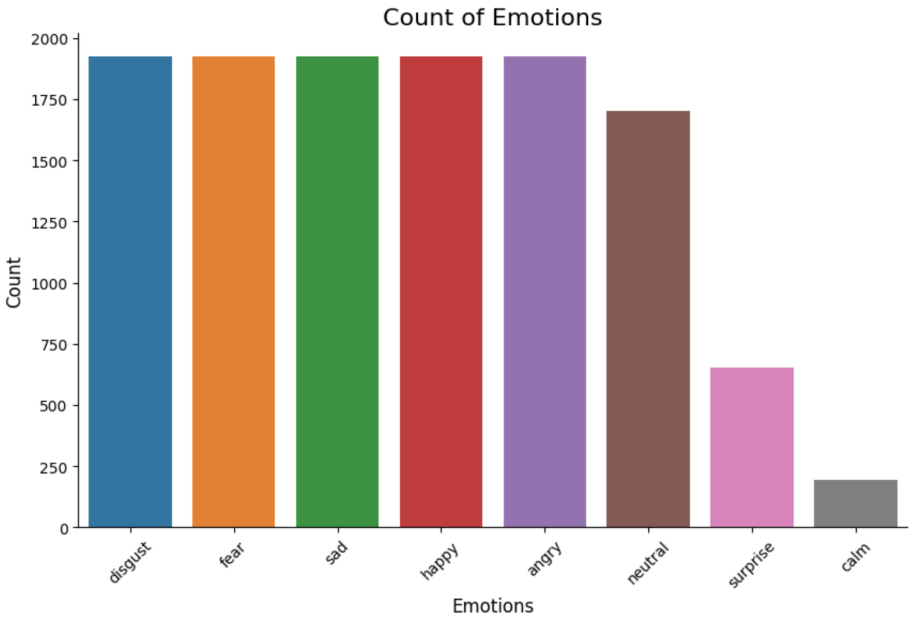


Figure 4.8: Visualise datasets emotions

4.2.2.4 Feature extraction

The audio signal is a three-dimensional signal in which three axes represent time, amplitude and frequency.

Some examples of features that can be extracted:

- Zero Crossing Rate : The rate of sign-changes of the signal during the duration of a particular frame.
- Energy : The sum of squares of the signal values, normalized by the respective frame length.

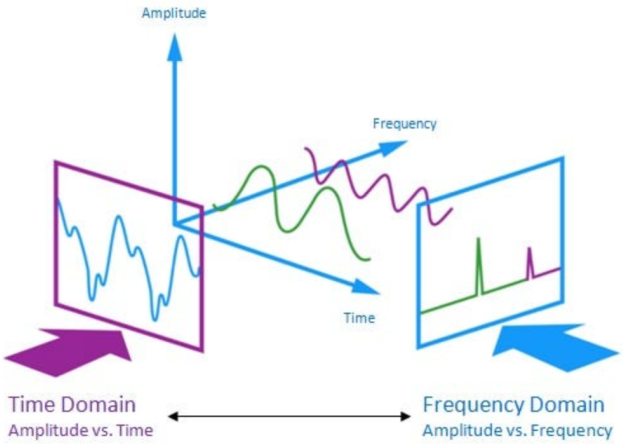


Figure 4.9: Feature extraction

- Entropy of Energy : The entropy of sub-frames normalized energies. It can be interpreted as a measure of abrupt changes.
- Spectral Centroid : The center of gravity of the spectrum.
- Spectral Spread : The second central moment of the spectrum.
- Spectral Entropy : Entropy of the normalized spectral energies for a set of sub-frames.
- Spectral Flux : The squared difference between the normalized magnitudes of the spectra of the two successive frames.
- Spectral Rolloff : The frequency below which 90% of the magnitude distribution of the spectrum is concentrated.
- MFCCs Mel Frequency Cepstral Coefficients form a cepstral representation where the frequency bands are not linear but distributed according to the mel-scale.
- Chroma Vector : A 12-element representation of the spectral energy where the bins represent the 12 equal-tempered pitch classes of western-type music (semitone spacing).
- Chroma Deviation : The standard deviation of the 12 chroma coefficients.

For the MVP we will extract the following features:

- Zero Crossing Rate
- Chroma stft
- MFCC
- RMS(root mean square) value
- MelSpectogram to train our model.

4.2.3 Model

We got a 62% precision which is about 20% lower than the most performant model. This Convolutional Neural Network (CNN) model is designed for audio speech emotion recognition, prioritizing simplicity for an exploration of accuracy. The architecture includes three convolutional layers with decreasing filter sizes, followed by max-pooling for feature extraction. Dropout is applied to mitigate overfitting. The model is then flattened and connected to two dense layers for classification. The Adam optimizer,

categorical crossentropy loss, and accuracy metric are used for compilation. This approach provides a concise yet effective foundation for understanding the interplay between model simplicity and accuracy in emotion recognition.

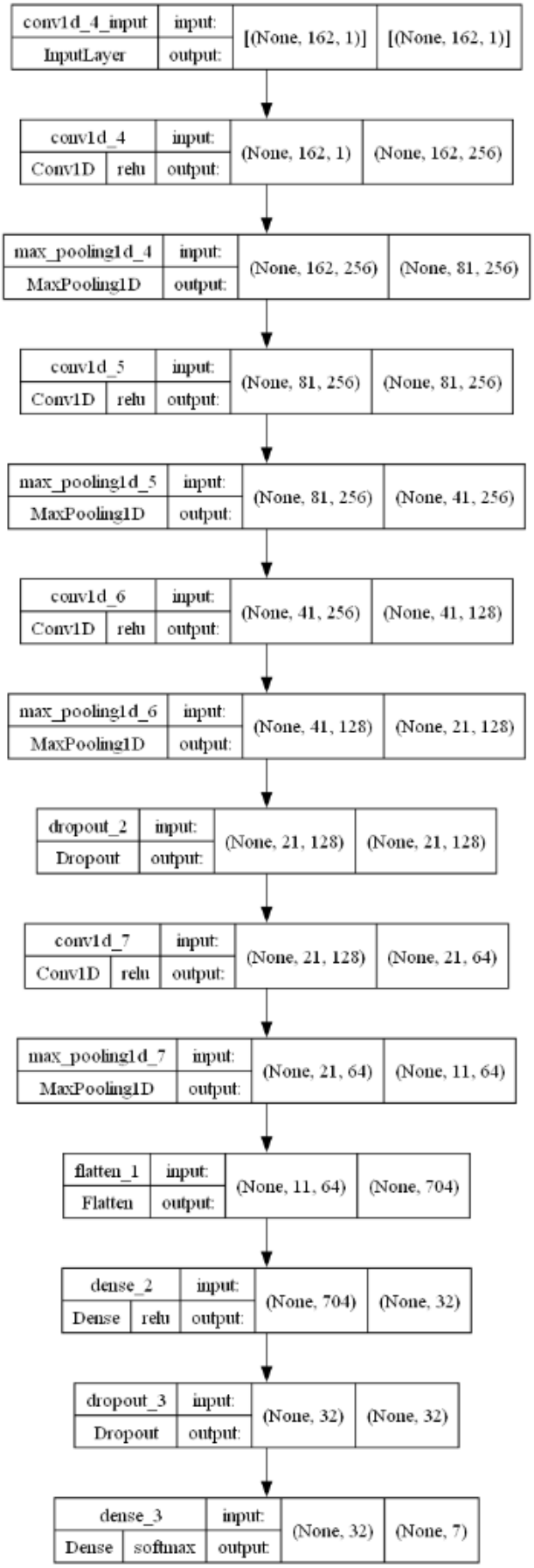


Figure 4.10: Model architecture

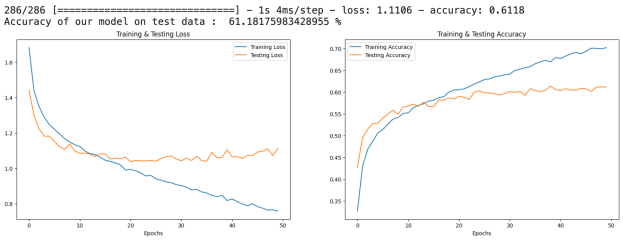


Figure 4.11: Accuracy

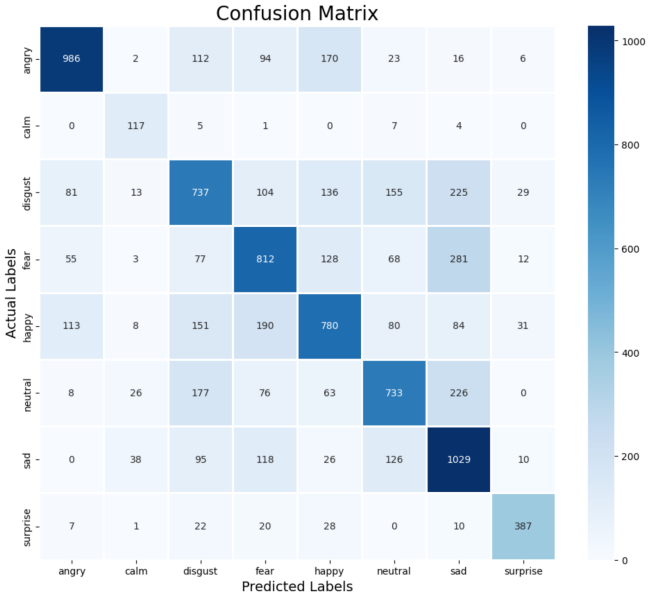
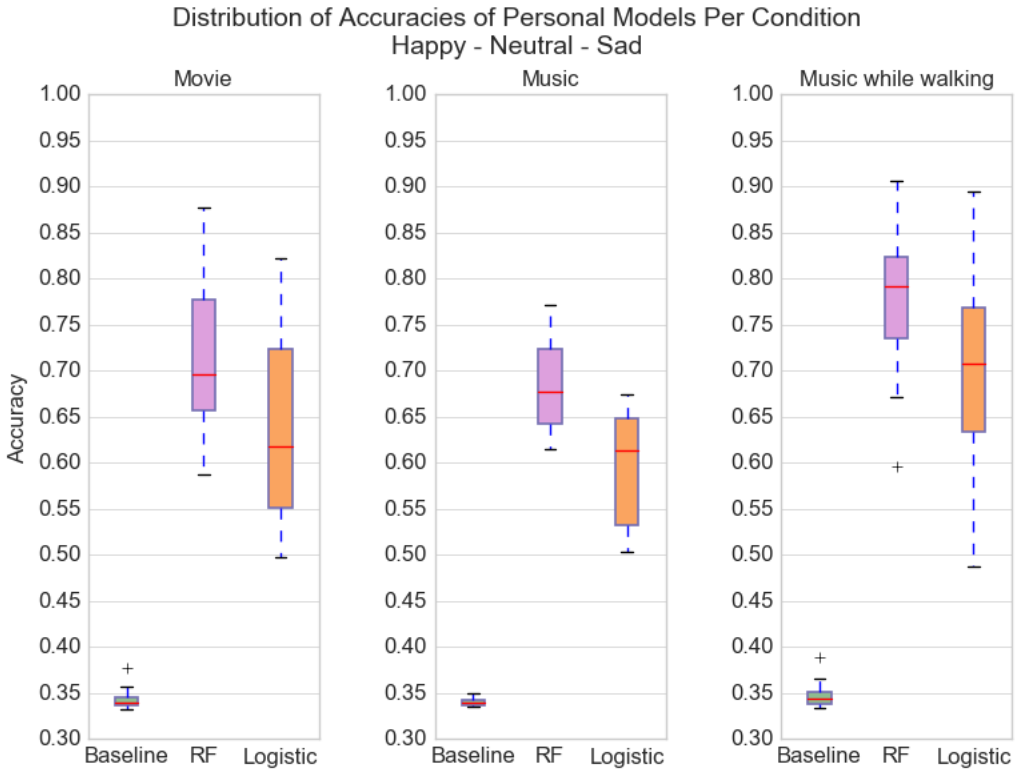


Figure 4.12: Accuracy

4.3 Emotion recognition based on physiological signals

We tried two different methods to classify emotions from physiological data. The first one used the heart rate and the accelerometer [8]data from 50 participants which were presented with happy, sad, and neutral stimuli(audio and audio-visual). Our results show that personal models outperformed personal baselines and achieved median accuracy higher than 78% for all conditions of the design study for binary classification of happiness versus sadness when the models are calibrated on the user. We used random forests and logistic regression as classifiers.

The benefits of this method are that we do not need complicated gadgets to collect data from participants; a simple wearable device is enough. The main disadvantage is that our models perform poorly when the user is not moving. Additionally, the model must be trained (calibrated) on some user data, which means that users must react to some stimuli and label their emotions appropriately for several minutes in order to achieve high accuracy. As a result, this model is limited to a few narrow usecases.

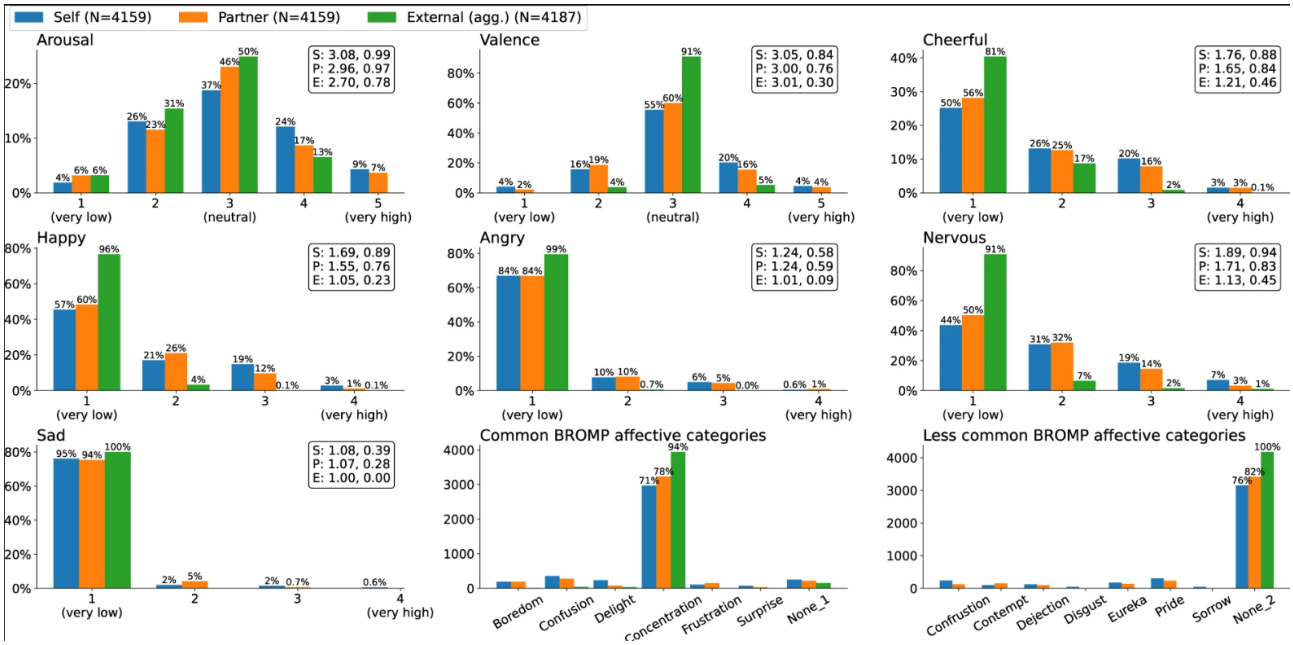


The second method used more data from the werable device, such as: BVP, EDA, ECG, TEMP. We used eXtreme Gradient Boosting(XGB) to clasify data into 4 labels high/low arousal and high/low valance. The data[9] was acquired from 16 sessions of approximately 10-minute long paired debates on a social issue. Our model achieved a median accuracy of 91% on these 16 sessions with the lowest being 77%.

4.3.1 Datasets

K-EmoCon[9] was designed in consideration of a social interaction scenario involving two people and wearable devices capable of unobtrusive tracking of physiological signals. The resulting K-EmoCon dataset contains multimodal data from 16 paired-debates on a social issue, which sum to 172.92 minutes of dyadic interaction. It includes physiological signals measured with three wearable devices, audiovisual recordings of debates, and continuous annotations of emotions from three distinct perspectives of the subject, the partner, and the external observers.

Annotations in our dataset for scaled emotions are highly biased. However, while arousal and valence are explicitly centered at zero (which corresponds to 3 = neutral), five emotions measured in the scale of 1 = very low to 4 = very high (cheerful, happy, angry, nervous, and sad) are systematically biased without a zero neutral. All of their values indicate that some emotion is present, and this absence of zero results in a widely varying interpretation of scale values by our participants and raters.



During a debate, participants wore a suite of wearable sensors, as shown in Fig. 2, which includes:

1. Empatica E4 Wristband â captured photoplethysmography (PPG), 3-axis acceleration, body temperature, and electrodermal activity (EDA). Heart rate and the inter-beat interval (IBI) were derived from Blood Volume Pulse (BVP) measured by a PPG sensor.

2 LookNTell Head-Mounted Camera â with a camera attached at one end of a plastic circlet, was worn on participantsâ heads to capture videos from a first-person POV.

4.3.2 Data preparation

Physiological signals were clipped from the respective beginnings of data collection sessions to the respective ends of debates, as the initial 1.5 to 2 minutes immediately after a session begins corresponds to a baseline measurement for a neutral state. For training the model the next features were extracted:

bvp: mean, hrv, mean-ibi, mse

gsr: eaks-per-sec, mean-amp, mean-risetime, mean-gsr, std-gsr

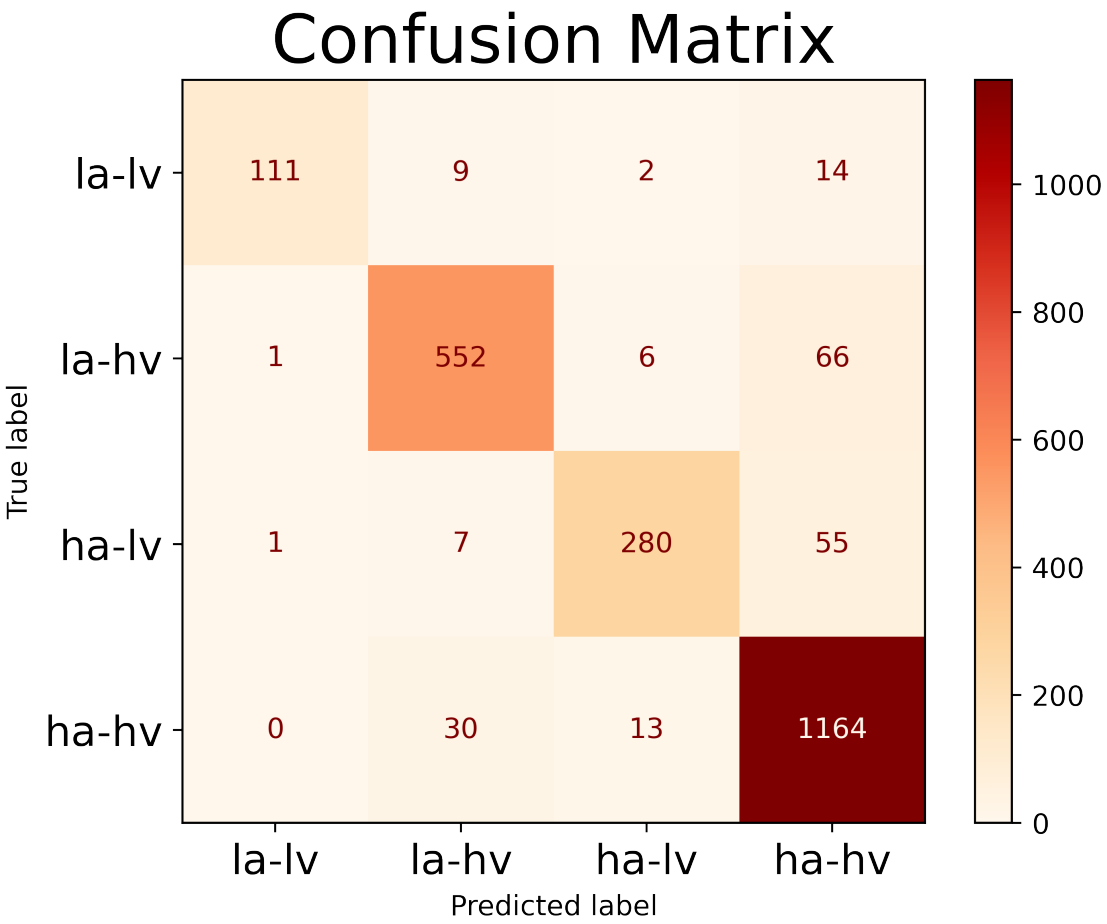
hst: mean, std, kurtosis, skew

ecg: mean, std

4.3.3 Validation

An experiment was conducted to investigate the influence of neighborhood bias on model evaluation using random cross-validation. In this experiment, a k-fold cross-validation was performed for each personal model, with the testing fold in each iteration containing either a contiguous happy data block or a contiguous sad data block. Used Leave-One-Out Cross-Validation to see how well does the model learn patterns from other users and with what accuracy can predict on a different user.

4.3.4 Results



4.3.5 Conclusion

We improved the accuracy made by our ML model to an mean accuracy of 90% without calibrating the model on the user.

Chapter 5

Application (Study case)

Main Functionalities and Specifications

Comprehensive Authentication Suite: This feature encompasses a full range of user authentication services, including:

- **User Sign-In:** Secure access to user accounts with robust verification processes.
- **User Sign-Up:** Streamlined registration procedure enabling new users to create an account effortlessly.
- **Password Management:** Allows users to change or reset their passwords, ensuring account security.
- **User Sign-Out:** A reliable sign-out mechanism that ensures user sessions are securely terminated.

Assessment History Tracking: This function enables users to:

- View a comprehensive history of all assessments conducted by them, including dates, scores, and other relevant details.
- Access and revisit previous assessments for review or comparison purposes.

Interactive Chart and Video Analytics: This utility offers:

- Dynamic charts that visually represent assessment data, enabling intuitive understanding and analysis.

- Video playback features that allow users to view their assessment videos with interactive elements for an immersive experience.

New Assessment Creation: Users can create new assessments, which involves the integration of the following advanced functionalities:

- Speech Emotion Recognition
- Image Emotion Recognition
- Psychological Signal Emotion Recognition

Speech Emotion Recognition

- **Functionality:** Analyzing emotional content within audio file.
- **Specification:** The application should run the deep learning model (that decodes emotions from speech, considering parameters such as pitch, tone, and speech rate) and create a json file and a plot showing at a specified period of time the emotion from the video.

Image Emotion Recognition

- **Functionality:** Recognizing emotions through facial expressions in the provided video.
- **Specification:** The application should run the deep learning model and create a json file and a plot showing at a specified period of time the emotion from the video. The model picks a specific number of frames from each window of time and analyzes them.

Psychological Signal Emotion Recognition

- **Functionality:** Integrating psychological signals, such as physiological data, into emotion recognition.
- **Specification:** The application accommodates CSV input for psychological signals, utilizing advanced algorithms to extract emotional insights from data such as pulse and other biofeedback from wearable devices.

5.1 Implementation

The application is a RESTful API, web client and server. The frontend is implemented in Angular, using several open-source libraries like MaterialUI, FontAwesome, D3, NgxCharts, AngularDomAPI. Meanwhile, the backend is coded in Java 17, it uses Spring Boot as the base framework to create a RESTful architecture and all implementations respect the REST principles.

Communication over https is secured and up-to-date with the latest standards, each (non-authentication) request bearing a JWT token encrypted with SHA256.

The models are exported and run with a python script as well as the algorithm that aligns all 3 models.

5.2 Testing

For testing the server, we employed unit tests, verifying each method and reaching a coverage of 96%. These tests cover every flow from the API and they combine seamlessly with the robust logging, present at every level in our server. For all of this we used JUnit5, the standard library for unit testing in Java. If the application is going to be further developed, we plan to create performance and stress tests, since the running of 3 AI models on a small server may cause unwanted performance bottlenecks and we need to be aware of it.

For ensuring the compatibility of the backend with the models, we created a few integration tests. These tests proved to be very useful, as they facilitated the parallel work on models and backend while ensuring the API between them is respected.

We only manually tested the frontend. Though very time intensive, we thoroughly tested it so that the rich user experience isn't compromised.

Chapter 6

Conclusion and future work

In conclusion, our project represents a significant venture into the realm of emotion recognition, where our primary focus has been on experimenting and focusing on these 3 main types of input: audio, video, signals from the digital watch. Despite the notable results, it is evident that there is still much work to be done. A noteworthy aspect of our project lies in the exploration of various models, and the potential for a multimodal approach that incorporates audio, video, and physiological signals.

Throughout this journey, we have successfully developed models that exhibit an accuracy surpassing 60%. Our journey started from understanding theoretical concepts in artificial intelligence to the practicalities of application development and working in a team.

In the future, engineers will need a larger and more varied data set in order to effectively correlate physiological data with audio and video emotions. A compelling avenue for further exploration involves the combination of speech-to-text and vocal emotion recognition (this intersection presents an intriguing opportunity to delve into more nuanced and comprehensive emotion analysis). As we continue to evolve our project, additional experiments in the direction of alignment the three main directions could yield valuable insights and contribute to the ongoing advancement of emotion recognition technologies.

Bibliography

[1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

[2] Muhammad Awais, Mohsin Raza, Nishant Singh, Kiran Bashir, Umar Manzoor, Saif Ul Islam, and Joel JPC Rodrigues. Lstm-based emotion detection using physiological signals: Iot framework for healthcare and distance learning in covid-19. *IEEE Internet of Things Journal*, 8(23):16863–16871, 2020.

[3] Muhammad Najam Dar, Muhammad Usman Akram, Sajid Gul Khawaja, and Amit N Pujari. Cnn and lstm-based emotion charting using physiological signals. *Sensors*, 20(16):4551, 2020.

[4] Yousif Khaireddin and Zhuofa Chen. Facial emotion recognition: State of the art performance on fer2013, 2021.

[5] Min Lin, Qiang Chen, and Shuicheng Yan. Network in network, 2014.

[6] Jia Le Ngwe, Kian Ming Lim, Chin Poo Lee, and Thian Song Ong. Patt-lite: Lightweight patch and attention mobilenet for challenging facial expression recognition. *arXiv preprint arXiv:2306.09626*, 2023.

[7] Karan Sharma, Claudio Castellini, Egon L van den Broek, Alin Albu-Schaeffer, and Friedhelm Schwenker. A dataset of continuous affect annotations and physiological signals for emotion analysis. *Scientific data*, 6(1):196, 2019.

[8] Giuseppe Romano Tizzano, Matteo Spezialetti, and Silvia Rossi. A deep learning approach for mood recognition from wearable data. In *2020 IEEE international symposium on medical measurements and applications (MeMeA)*, pages 1–5. IEEE, 2020.