UNIVERSITÉ DU QUÉBEC EN OUTAOUAIS

MULTISENSOR AFFECT RECOGNITION FOR DATA COLLECTED DURING VIRTUAL REALITY IMMERSIONS

THÈSE PRÉSENTÉE

COMME EXIGENCE PARTIELLE

DU PROGRAMME DE DOCTORAT EN SCIENCES ET TECHNOLOGIES DE

L'INFORMATION

PAR

Itaf Omar Joudeh

Août 2025

EVALUATION JURY

Président du jury :	Dre. Nadia Baaziz
Membre interne du jury :	Dr. Etienne St-Onge
Membre externe du jury :	Dr. Yannick Francillette
Directeur de recherche :	Dre. Ana-Maria Cretu
Codirecteur de recherche :	Dr. Stéphane Bouchard

DEDICATION

I sincerely dedicate this thesis to my precious parents and siblings, and most importantly, to my one and only beloved husband. They have always been there for me and stood by my side throughout this journey. They are the ones who keep me cheerful the whole time, and push me forward in this life. *"I adore you ALL!"*

ACKNOWLEDGEMENTS

I would like to express my gratitude to my supervisor, Dr. Cretu. She has been a great mentor, who directed me in the right direction during the toughest times. She has always treated me with kindness and respect. *"Thank you for being a growth friend before being a supervisor!"*

I would like to acknowledge Dr. Bouchard for his time and help. He has provided me with valuable insight and resources in relation to cyberpsychology and virtual reality (VR). He has also supplied a data set collected at the Laboratoire de cyberpsychologie of the Université du Québec en Outaouais (UQO). *"This is much appreciated!"*

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

TABLE OF ABBREVIATIONS, INITIALS, AND ACRONYMS

Abbreviation/Acronym	Description
2D	Two-Dimensional
3D	Three-Dimensional
AFEW	Acted Facial Expressions in the Wild
Aff-Wild	Affect in the Wild
AI	Artificial Intelligence
AR	Affect Recognition
AU	Action Unit
AVEC	Audio/Visual Emotion Challenge
AVRS	Affective Virtual Reality System
BARGAIN	Behavioral Affective Rule-based Games Adaptation Interface
BiLSTM	Bidirectional Long Short-Term Memory
BoAW	Bags-of-Audio-Words
BR	Breathing Rate
CCC	Concordance Correlation Coefficient
CNN	Convolutional Neural Network
ComParE	Computational Paralinguistics ChallengE
CV	Coefficient of Variation
DAG	Directed Acyclic Graph
ECG	Electrocardiogram
EDA	Electrodermal Activity
EEG	Electroencephalography
eGeMAPS	Extended Geneva Minimalistic Acoustic Parameter Set
EMG	Electromyography
EOG	Electrooculography
ESN	Echo State Network
FACS	Facial Action Coding System
FFT	Fast Fourier Transform
FSM	Finite State Machine
GeMAPS	Geneva Minimalistic Acoustic Parameter Set
GEQ	Game Experience Questionnaire
GES	Gold-standard Emotion Sub-challenge
GPU	Graphics Processing Unit
GSR	Galvanic Skin Response
HMD	Head-Mounted Display
HR	Heart Rate
HRV	Heart Rate Variability
HTC	High Tech Computer
IEEE	Institute of Electrical and Electronics Engineers
IRB	Institutional Review Board

Abbreviation/Acronym	Description
IVE	Immersive Virtual Environment
LFPC	Log Frequency Power Coefficient
LGBP-TOP	Local Gabor Binary Patterns from Three Orthogonal Planes
LLD	Low-Level Descriptor
LSBoost	Least-Squares Boosting
LSTM	Long Short-Term Memory
MDPI	Multidisciplinary Digital Publishing Institute
MFCC	Mel-Frequency Cepstral Coefficients
MRMR	Minimum Redundancy Maximum Relevance
MSE	Mean Squared Error
N/A	Not Applicable
NaN	Not a Number
NCA	Neighborhood Component Analysis
NLD	Normalized Length Density
NN	Neural Network
NPC	Non-Player Character
NSI	Non-Stationary Index
PCC	Pearson Correlation Coefficient
POC	Proof-of-Concept
PPG	Photoplethysmography
RECOLA	Remote Collaborative and Affective Interactions
REB	Research Ethics Board
ResNet	Residual Network
RIP	Respiratory Inductive Plethysmography
RMSE	Root Mean Squared Error
RNN	Recurrent Neural Network
RSP	Respiration
SAM	Self-Assessment Manikin
SCL	Skin Conductance Level
SCR	Skin Conductance Response
SDC	Shifted Delta Cepstrum
SFC	Sum of Fixation Counts
SGDM	Stochastic Gradient Descent with Momentum
SKT	Skin Temperature
SPL	Saccade Path Length
STD	Standard Deviation
SVM	Support Vector Machine
SVR	Support Vector Regression
UQO	University of Quebec in Outaouais
VR	Virtual Reality

RÉSUMÉ

Cette étude vise, à long terme, à concevoir et à développer une nouvelle intervention adaptable pour traiter les troubles cognitifs en utilisant la réalité virtuelle (RV). L'état affectif d'une personne reflète son état émotionnel/cognitif, qui pourrait être mesuré par les dimensions d'activation et de valence, selon le modèle Circumplex. Cette thèse propose et valide plusieurs nouvelles solutions d'apprentissage automatique dans le but ultime d'ajuster de manière adaptative les environnements de RV aux états affectifs des utilisateurs. Les modèles d'apprentissage automatique proposés ont été utilisés pour la prédiction des valeurs d'activation/valence à partir de diverses sources de données, à savoir des données physiologiques, visuelles et acoustiques. Dans cette thèse, les valeurs d'activation/valence ont été initialement prédites en exploitant les enregistrements de l'ensemble de données Remote Collaborative and Affective Interactions (RECOLA), afin d'identifier les modèles les plus prometteurs à ajuster, tester et valider à l'aide de données recueillies lors d'immersions en RV au Laboratoire de cyberpsychologie de l'UQO. Les signaux d'activité électrodermale (EDA) et d'électrocardiogramme (ECG) de RECOLA représentent des données physiologiques, tandis que les enregistrements vidéo et audio de RECOLA représentent des données visuelles et acoustiques pour l'expérimentation dans cette thèse. Divers schémas de fusion de décision ont également été appliqués pour combiner les prédictions d'activation et de valence des régresseurs les plus performants. Une régression multimodale a également été réalisée pour prédire les valeurs d'activation/valence à partir de caractéristiques physiologiques, visuelles et acoustiques combinées. La fusion des décisions et des caractéristiques a permis d'améliorer les performances de prédiction.

Les principales contributions de la thèse comprennent un test empirique auprès de personnes immergées en RV, basé sur des données recueillies au Laboratoire de cyberpsychologie de l'UQO à cette fin. Des données physiologiques (caractéristiques de l'ECG) et visuelles (caractéristiques des yeux et du visage) ont été recueillies lors des immersions en RV. Les données collectées ont ensuite été synchronisées, traitées et étiquetées avec les annotations d'activation/valence correspondantes. Des régresseurs d'apprentissage automatique ont ensuite été testés pour prédire les valeurs d'activation/valence. Bien qu'une comparaison directe n'ait pas été possible, les performances obtenues sont cohérentes avec celles disponibles dans la littérature pour la prédiction d'activation/valence. Sur la base des résultats de l'analyse réalisée sur les données RECOLA et sur celles collectées lors de l'étude de RV, les solutions proposées ont démontré leur potentiel d'intégration dans un système de RV pour la réadaptation cognitive des patients souffrant de schizophrénie. L'objectif ultime est d'optimiser automatiquement le niveau de charge mentale demandé par les utilisateurs de RV. Cela peut faciliter les exercices de remédiation cognitive pour les utilisateurs souffrant de troubles de santé mentale, tout en évitant le découragement. Enfin, les contributions de la thèse ont le potentiel de transformer les traitements proposés et d'avoir un impact sur la vie des patients en réduisant les déficiences cognitives liées à la schizophrénie et à d'autres troubles mentaux.

ABSTRACT

This study is aimed, in the long term, at the design and development of a novel adaptable intervention to treat cognitive impairments using virtual reality (VR). The affective state of a person reflects their emotional/cognitive state, which could be measured through the arousal and valence dimensions, as per the Circumplex model. This thesis proposes and validates multiple novel machine learning solutions with an eventual aim to adaptively adjust VR environments to the affective states of users. The proposed machine learning models were used for the prediction of arousal/valence values from various data sources, namely physiological, visual, and acoustic data. In this thesis, arousal/valence values were initially predicted by exploiting recordings from the Remote Collaborative and Affective Interactions (RECOLA) data set, in order to identify the most promising models to be adjusted, tested, and validated using data collected from immersions in VR at the Cyberpsychology Laboratory of UQO. The electrodermal activity (EDA) and electrocardiogram (ECG) signals of RECOLA represent physiological data, whereas the video and audio recordings of RECOLA represent visual and acoustic data for experimentation in this thesis. Various schemes of decision fusion have also been applied to combine the arousal and valence predictions of the best performing regressors. Multimodal regression was also performed to predict arousal/valence values from combined physiological, visual, and acoustic features. Decision and feature fusion led to better prediction performances.

The main contributions of the thesis include an empirical test with people immersed in VR, based on data that were collected at the Cyberpsychology Laboratory of UQO for this purpose. Physiological (ECG features) and visual (eye and facial features) data

were collected during VR immersions. The collected data were then synchronized, processed, and labelled with the corresponding arousal/valence annotations. Machine learning regressors were then tested to predict arousal/valence values. While a direct comparison was not possible, the reached performances are coherent with the ones available in the literature for arousal/valence prediction. Based on the results of the analysis carried out on the RECOLA data set and on the data set collected during the VR study, the proposed solutions demonstrated their potential for integration into a VR system for cognitive rehabilitation of patients suffering from schizophrenia. The ultimate goal is to automatically optimize the level of cognitive effort requested by VR users. This can help facilitate cognitive remediation exercises for users with mental health disorders, while avoiding discouragement. Finally, the contributions in the thesis have the potential to transform the treatments offered and have an impact on patients' lives by reducing cognitive impairments linked to schizophrenia and other mental disorders.

1. INTRODUCTION

This chapter provides the motivation behind the thesis, describes the problem statement, reviews the objectives and contributions, and outlines the structure of the thesis.

1.1. Motivation

Affect recognition (AR) is a signal and pattern recognition process that plays a major role in affective computing [1]. Affective computing inspires the development of devices that are capable of detecting, processing, and interpreting human affective states. As such, AR is an interdisciplinary research area which includes signal processing, machine learning, psychology, and neuroscience. The affective state refers to the emotional/cognitive condition or mood of an individual at a given time [1,2]. The prediction of affective states can help researchers in a variety of fields. For example, various systems can be optimized based on the affective state of the user for a better experience. AR can also aid psychologists in the diagnosis of mental health disorders. AR can detect the affective state of a person by monitoring their activity and vital signs through sensors. AR can then classify affective states by analyzing physiological data, acoustic data, and/or visual data [3,4]. Integrating AR into virtual reality (VR) can unlock great potential.

VR is a mean of visualization that minimizes outside distractions. VR systems usually use head-mounted displays (HMD), covering the eyes of users to display images that are rendered in real time and adapted to the user's movement via built-in motion trackers [5]. VR restrains visual information to what is displayed in the HMD and

applies interactivity methods to create a sense of presence, making users feel like they are actually in the presented environment. The immersion of senses benefits many fields, which can range from entertainment and education to psychological therapy. Interactivity in VR creates a connection between the application and the user. VR interprets movements from the physical world to the virtual world in a precise manner. This can be designed via controls that simulate physical actions, allowing the user to attain positional awareness easily. One can benefit from this paradigm by virtually enabling visual representations of workflows and data. VR implements intuitive movements for interaction, in an environment which has the ability to represent three-dimensional (3D) objects. VR provides a method of visualization that has the potential for mimicking day-to-day life in order to learn complex skills.

Technological advancements have made VR viable in a wide range of clinical applications [5]. One promising application is cognitive intervention for people with mental disorders, where AR could help in personalizing the level of difficulty of training tasks to each individual, while avoiding discouragement. A person can be monitored using sensors such as electrodermal activity (EDA), electrocardiography (ECG), and an external camera, while doing various tasks in VR. Physiological data collected by the sensors and facial, visual data collected by the camera can then be used to determine the affective state of that person.

1.2. Problem Statement

A novel machine learning approach is required to seamlessly adjust a VR environment to match the optimal affective state of individuals who are engaged in cognitive interventions, while immersed in VR and their face partially covered by the VR HMD. Machine learning models are capable to learn from data from several people to better adapt the VR environment. Such a solution can aid doctors in their diagnostic

operations, for example, by triggering an alarm in the case of distress for the individual, and thus, allowing them to intervene promptly in the treatment. Future studies could explore the use of additional sensors to not only predict affective states, but also measure cognitive effort during VR interventions, enabling the development of more personalized and effective treatments for individuals with cognitive impairments. Meaningful means can also be proposed to use the measured affective state to adapt the feedback to support users in specific tasks (i.e., control the volume of information displayed using selectively-densified objects, provide additional task support, for example, by restricting the number of possible movements in 3D to the movements that are necessary in the task or by restricting to one type of movement or one direction at a time, offering support in unexpected conditions, etc.).

1.3. Objectives and Expected Contributions

This work represents an initial step towards predicting affective states through physiological, visual, and acoustic data that will be obtained from VR immersions. The ultimate goal is to design, develop, implement, and validate an adaptable VR system that can aid in the treatment of cognitive impairments in individuals with mental health disorders, through a synergistic approach that blends computer science and psychological techniques. The context of this thesis covers the machine learning aspect, where the affective states of emotions are predicted based on arousal and valence values.

This thesis focuses on developing novel machine learning solutions based on visual (image/video), acoustic (audio recordings), and physiological sensory data (EDA and ECG) to help adaptively adjust a VR environment to the user's affective state in future work, beyond the scope of this thesis. Thus, AR was performed as a proof-of-concept to predict the user's affective state using the Remote Collaborative and Affective

Interactions (RECOLA) data set. Then, a preliminary test of AR was performed for a pilot VR study. The thesis was conducted in several stages of research:

1. Leveraging and analyzing available data sets to find a data set that contains suitable data for AR;
2. Extraction, processing, and fusion of physiological, visual, and acoustic features;
3. Training, validating, and testing classical machine learning solutions such as regression trees and ensemble regressors to perform AR on physiological data, including sample EDA and ECG readings and their features;
4. Training, validating, and testing deep machine learning solutions such as a recurrent neural network (RNN) to perform AR on physiological (EDA and ECG) sequences;
5. Exploring multiple processing techniques to boost the prediction performance of the regressors;
6. Training, validating, and testing classical machine learning solutions such as regression trees, kernel regression, and ensemble regressors to perform AR on visual features;
7. Applying video and image processing techniques to prepare visual data for AR;
8. Training, validating, and testing deep machine learning solutions such as pretrained ResNet-18 and MobileNet-v2 convolutional neural networks (CNNs) to perform AR on face images;
9. Experimenting with the training parameters and options of the CNNs to enhance their prediction;
10. Using the trained CNNs to extract deep facial features that can be inputted into classical machine learning regressors in order to perform AR;
11. Training, validating, and testing classical machine learning solutions such as regression trees and ensemble regressors to perform AR on acoustic features;

12. Feature fusion and training, validating, and testing classical machine learning solutions such as regression trees and ensemble regressors to perform AR on multimodal features (i.e., combined physiological, visual, and acoustic features);
13. The analysis, visualization, and comparison of predictions obtained from multiple solutions to arrive at the best solution for each data modality;
14. Extracting and preprocessing physiological and visual data that were collected in a VR setting;
15. Data analysis of the data collected during VR immersions;
16. Synchronization, processing, and labelling of the data collected during VR immersions;
17. Feature fusion of physiological and visual data collected during VR immersions;
18. Performing a preliminary test on the collected data to perform AR via machine learning; and
19. Generation of data samples to create more data, and add variation between data samples in order to enhance the prediction performance.

Since this project is part of a larger interdisciplinary initiative aimed at improving the current capabilities for cognitive rehabilitation of people with schizophrenia, the most promising models are tested on a VR system designed for this purpose in the Cyberpsychology Laboratory of UQO, as later described in this thesis. Ultimately, the aim is to automatically optimize the level of cognitive effort requested during VR-based cognitive remediation exercises to avoid the discouragement of participants with schizophrenia. These aspects ensure the originality of this Ph.D. project. Furthermore, this research is one step towards the movement into a metaverse society, which ensures inclusion for everyone around the world, regardless of their health or economic status (i.e., VR for all).

The expected contributions of this thesis are as follows:

1. Proposing a novel combination of processing steps for different data modalities;
2. Extracting deep facial features for AR;
3. Introducing an innovative combination of features;
4. Proposing an outliers-based feature selection mechanism;
5. Introducing a unimodal and multimodal decision fusion mechanism;
6. Performing feature fusion and multimodal AR;
7. Submitting a research protocol and obtaining approval from a Research Ethics Board (REB) or Institutional Review Board (IRB);
8. Extracting and preprocessing physiological and visual data collected during VR immersions; and
9. Applying AR on the data from VR immersions.

1.4. Thesis Structure

This thesis is organized into ten chapters. The next two chapters provide a background about the topic (Chapter 2) and outline the proposed solution (Chapter 3), respectively. The remainder of the thesis discusses the work carried out to achieve the research stages and contributions mentioned in Section 1.3. As such, Chapters 4 and 5 present the proposed solutions and results on physiological and visual data, respectively. Chapter 6 introduces a proof-of-concept solution that can be applied for visual data collected through a VR system at the final stage of the thesis. Chapter 7 presents the proposed solution and results on acoustic data. Chapter 8 presents the proposed solution for multimodal AR with combined physiological, visual, and acoustic features. Chapter 9 presents the results for physiological and visual data collected through VR immersion, during a pilot study in the Cyberpsychology Laboratory of UQO. It also details the conducted study involving data collection experiments, data processing, an analysis of the data, and the application of AR. Finally, Chapter 10 concludes the thesis.

2. BACKGROUND AND LITERATURE REVIEW

This chapter describes some of the existing emotional models in the literature (Section 2.1). It also provides a brief background about artificial intelligence (AI) in relation to AR (Section 2.2). This chapter then discusses common solutions for AR problems, available data sets used in AR applications, and state-of-the-art AR studies (Section 2.3). Lastly, this chapter includes an overview about VR (Section 2.4), and the application of AR within VR in the literature (Section 2.5).

2.1. Emotional Models of Affective States

Researchers have developed emotional models to describe humans' affective states [1,3]. There are two types of emotional models: categorical and dimensional. In categorical models, discrete categories are used to express different emotions. In dimensional models, "emotions are mapped into a multidimensional space, where each of the axis represents a continuous variable." The Circumplex model is one example of dimensional models [1,6].

The Circumplex model [6] (see Figure 1) is widely used in AR studies [7-10] to predict affective states. In this dimensional model, affective states are characterized as discrete points in a two-dimensional (2D) space of valence and arousal axes. Arousal is used to rate the activity/energy level of the affective state. Valence is used to rate the positivity of the affective state. There are four quadrants in the Circumplex model: low arousal/low valence, low arousal/high valence, high arousal/low valence, and high arousal/high valence. The four quadrants are attributed with the sad, relaxed, angry,

and happy affective states, respectively. When both arousal and valence are low, the resulting emotion is typically sadness. When arousal is low and valence is high, the person tends to feel relaxed. Conversely, high arousal and low valence often correspond to feelings of anger or anxiety. Finally, when both arousal and valence are high, it usually indicates a state of happiness.

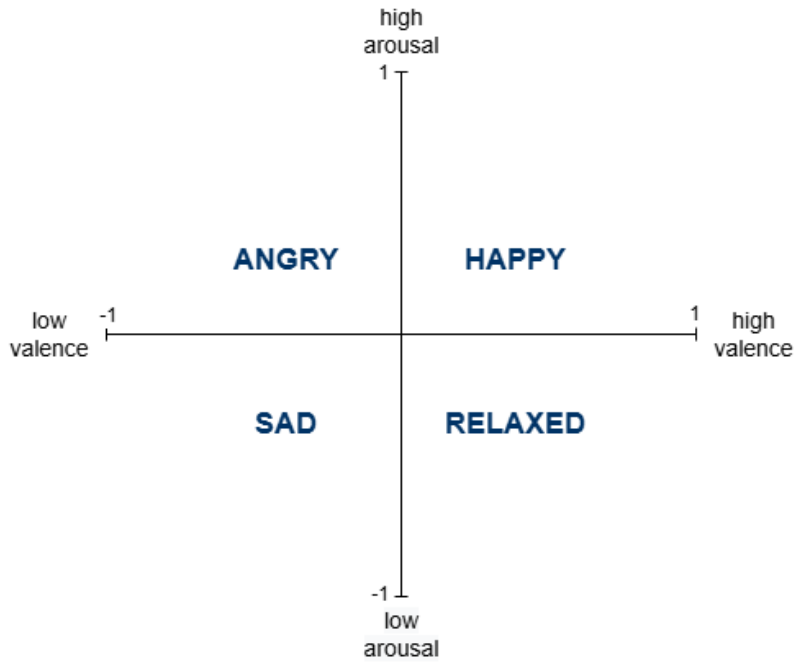


Figure 1. The Circumplex model of affective states [6].

2.2. Artificial Intelligence (AI)

AI is a field of study that is aimed at the development of intelligent machines. AR relies on AI solutions such as machine learning. AR is the process of detecting the affective state of a person by monitoring his/her activity and vital signs through sensors [1-3]. The affective state of a person can be described through continuous emotional predictions of arousal and valence values. AI-based techniques can detect/recognize affective states by analyzing physiological data, human voice, gaze, facial expressions,

body gestures, and standard interactions like mouse movements and keystrokes [3,4]. Hence, AR can be based on physiological data from wearable sensors, and/or on audio-visual data from a camera. As machine learning was used in order to develop the solution for continuous emotional predictions of arousal and valence values, this section presents a brief description of machine learning solutions that can be used in the context of AR.

2.2.1. Machine Learning

As suggested by the name, machine learning is the process where a machine uses a set of training data to learn how to perform a task [11,12]. There are two types of machine learning: supervised or unsupervised. In supervised learning, information about the expected output is required alongside the training data. In unsupervised learning, the machine depends on the patterns associated with the training data to determine the output. In this study, supervised machine learning models were leveraged.

Supervised machine learning methods consist of a classifier or regressor which is trained on a data set of labelled attributes/features to fit into the model of the classifier or regressor [11,12]. The trained classifier or regressor can then be exploited to identify new, unlabelled data samples by accordingly predicting their labels. Classifier/regressor models can be linear, quadratic, cubic, or Gaussian in nature. In a classification problem, a classifier is trained to classify/group data samples into categorized classes. In a regression problem, a regressor is trained to estimate the labelled numerical continuous-response values of data samples. In this study, AR was executed as a regression problem to predict the arousal and valence values of affective states.

Examples of supervised learning algorithms include decision trees, support vector machines (SVMs), and ensemble models among many others [11]. The branches of a

decision tree represent condition-decision pairs that are dependent on the boundaries of the attributes in the training data set. SVMs try to find the ideal hyperplane for labelling data samples. An ideal hyperplane identifies a clear gap between the data samples in terms of space. Ensemble models combine multiple learners via bagging, random space, or a boosting algorithm to optimize the prediction performance by aggregating predictions from their learners [13].

The validation of the learning process is an important step as it helps to prevent the classifier/regressor from overfitting [11]. Overfitting happens when the learner memorizes the training data. It causes the trained model to fail at predicting new data samples, that were not provided during the training process. Class imbalance is another issue that is particularly common in classification problems. It occurs when a data set has a lot more samples of one class (majority class) than the other class (minority class). Solutions to class imbalance include discarding the excess samples of the majority class, obtaining more samples of the minority class(es), oversampling and undersampling, or synthesizing new data samples.

The training process can be validated by assessing the trained model against a testing/validation data set, which contains part of the original training data set that was not included in the training, or another data set that contains similar samples [11]. Holdout testing and k-fold cross validation are two validation methods that can be used in machine learning. Holdout testing divides the data into a training data set and a testing data set, where a larger amount of data is kept for training while the remaining data are held out for testing at a later stage. In classification problems, the different classes in the data are equally distributed among the training and testing data sets to avoid overfitting. Holdout testing is preferred for large training data sets with imbalanced classes. K-fold cross validation is performed by repetitively training the classifier/regressor on $k - 1$ folds (i.e., partitions) of the data, and testing it on the

remaining fold. It is an iterative testing process, where the average of the *k* iterations is used to determine the final decision. This constantly checks the ability of the classifier/regressor at identifying new, unseen data, and hence, protects against overfitting. In this study, 5-fold cross validation was used to protect the regressors against overfitting.

2.2.2. Deep Learning

Deep learning can also be used for AR. Deep learning is a higher level of machine learning that is broadly employed in many fields nowadays [5]. Deep learners consist of deep neural networks that have many neurons [5,13,14]. Deep learning models are constructed with multiple levels of abstraction, where a backpropagation algorithm evaluates the amount of adjustment in a machine's parameters between the different levels. They are potentially capable of analyzing the patterns associated with complex types of data such as images, videos, sequences, or time series. Hence, they unlock advanced applications like object recognition, speech recognition, face recognition, and activity recognition.

RNNs and CNNs represent a specific category of deep neural networks, which are designed for sequential and/or structured data [13]. Sequential data can come from signals, audio, video, and visual imagery. RNNs are deep learning structures that use past information to enhance the performance of the network on current and future inputs. RNNs contain hidden states and loops. Their looping structure allows them to store past information in the hidden state and operate on sequences. Applications of RNNs include voice recognition, and text recognition. In this study, RNNs were used to predict the arousal and valence values of affective states from sequences of EDA and ECG data.

CNNs have many layers that can learn to extract different features of images [13]. The layers apply filters to training images at different resolutions, where the outputted images are then used as the input to the next layer. The filters in initial layers start to detect simple features, such as brightness and edges. Eventually, the filters can increase in complexity to detect features that uniquely define an object. Applications of CNNs include the labelling of pathology images on a per-pixel basis [15], object recognition for self-driving cars [16], and segmentation in medical imaging [17]. In this study, CNNs were used to predict the arousal and valence values of affective states from images of human faces.

2.2.3. Prediction Performance Metrics

The performance of a regressor can then be measured by comparing the predicted values with the actual values of the data samples. There are different metrics that can be calculated to measure the performance of a regressor: root mean squared error (RMSE), Pearson's correlation coefficient (PCC), and concordance correlation coefficient (CCC).

The RMSE simply measures the root of the mean of squared difference between the set of arousal/valence predictions and the set of the actual values [13,18,19]. The RMSE is calculated as follows:

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{N}} \tag{1}$$

where N is the total number of samples; $\hat{y}_i$ and y_i are the predicted and actual values for a given sample at time/index, i; and n is the number of samples in a subset of data. A smaller RMSE value represents better performance.

PCC measures the linear correlation between the set of arousal/valence predictions and the set of the actual values [13,18,19]. PCC is estimated using the following formula:

$$\mathrm{PCC} = \rho = \frac{\sum_{i=1}^{n}(\hat{y}_i - \mu_{\hat{y}})(y_i - \mu_y)}{\sqrt{\sum_{i=1}^{n}(\hat{y}_i - \mu_{\hat{y}})^2 \cdot \sum_{i=1}^{n}(y_i - \mu_y)^2}} \quad (2)$$

where N is the total number of samples; $\hat{y}_i$ and y_i are the predicted and actual values for a given sample at time/index, i; n is the number of samples in a subset of data; and $\mu_{\hat{y}}$ and μ_y are the mean values of the predicted and actual sets.

The CCC measure combines PCC with the square difference between the mean of the set of arousal/valence predictions and the set of the actual values [13,18]. The CCC can be evaluated by the following equation:

$$\mathrm{CCC} = \rho_c = \frac{2\rho\sigma_{\hat{y}}\sigma_y}{{\sigma_{\hat{y}}}^2 + {\sigma_y}^2 + \left(\mu_{\hat{y}} - \mu_y\right)^2} \quad (3)$$

where ρ is PCC between the predicted and actual sets; $\sigma_{\hat{y}}$ and σ_y are the standard deviations (STDs) of the predicted and actual sets; and $\mu_{\hat{y}}$ and μ_y are the mean values of the predicted and actual sets. Larger PCC and CCC values indicate a better performance.

These measures will be employed in this thesis to evaluate the performance of the designed solution for arousal and valence predictions.

2.3. Affect Recognition (AR)

This section discusses common solutions for AR problems (Section 2.3.1), and describes data sets available for AR applications (Section 2.3.2). Finally, it provides a literature review on state-of-the-art AR studies (Section 2.3.3).

2.3.1. Solutions

Most state-of-the-art AR studies predict continuous emotional values such as the arousal and valence dimensions. They use wearable sensors for their low-cost, rich functionality, and valuable insights [1]. Wearable sensors are sensors that can be worn on the human body or inserted into clothing. Integrated wearable sensor applications can be used in daily lives as they are portable and non-intrusive. They can be used to measure physiological data such as electroencephalography (EEG), electrooculography (EOG), ECG, electromyography (EMG), respiratory inductive plethysmography (RIP), blood oxygen, blood pressure, photoplethysmography (PPG), temperature, EDA, inertia, position, voice, etc. Wearable-based AR systems can be used in various healthcare applications to monitor the affective states of individuals [1].

AR can also be based on audio-visual data, which depend on multimodal features. These features are extracted from images or video. Audio-visual AR systems that are based on images of faces ideally consist of two components: parametrization and recognition of facial expressions (classification/regression), where parameterization usually consists of three main stages: face localization, face registration, and feature extraction (predesigned or learned) [3]. Parametrization is the process of specifying the visual features and coding schemes to describe the involved facial expressions. The visual features used for the prediction of cognitive states can be appearance or geometric features [4]. Geometric features represent the geometry of the face. Local

Gabor Binary Patterns from Three Orthogonal Planes (LGBP-TOP) [20] is one method that is used in the extraction of appearance features, while facial landmarks [21] are usually used for geometric features. Examples of geometric features include the derivatives of the detected facial landmarks, the speed and direction of motion in facial expressions, the head pose, and the direction of the eye gaze. Appearance features represent the overall texture resulting from the deformation of the neutral facial expression. Appearance features depend on the intensity of an image, whereas geometrical features determine distances, deformations, curvatures, and other geometric properties [3]. Coding schemes can either be descriptive or judgmental. Descriptive coding schemes depend on surface properties and what the face can do to describe facial expression. Judgmental coding schemes depend on the latent emotions or affects that produce them to parameterize facial expressions. The facial action coding system (FACS) [22] is one example of descriptive coding systems. It is a system that describes all visually evident facial movements [22,23]. It divides facial expressions into individual components of muscle movement, called Action Units (AUs). Coding schemes such as facial AUs, as well as geometric and/or appearance features can then be treated as input parameters to machine learning regressors or classifiers for the prediction of cognitive states.

There are three common data modalities currently being considered for visual AR solutions: RGB, 3D, and thermal [3]. Multimodal fusion is an additional step that is usually performed to combine multiple data modalities. There are three types of multimodal fusion: early, late, and sequential fusion [3]. Early fusion combines the modalities at the feature level, while late fusion combines the modalities at the decision level. Sequential fusion combines different modality predictions sequentially. Combining modalities at the different levels provides more information to machine learners and can potentially help to achieve better prediction performances.

2.3.2. Data Sets

The application of AR requires a large amount of data, collected from a diverse group of participants. Researchers have published data sets to enable the validation and comparison of results. Data sets can be grouped based on content, data modality, and/or participants [3]. Such data sets are composed of posed or spontaneous facial expressions, primary expressions, or facial action units as labels, still images or video sequences (i.e., static/dynamic data), and controlled laboratory or uncontrolled non-laboratory environments. Example data sets include, but are not limited to RECOLA, AFEW (Acted Facial Expressions in the Wild), Aff-Wild (Affect in the Wild), and SEWA [2].

As part of research stage 1 (see Section 1.3), the RECOLA data set was chosen since it includes annotated recordings for multiple data modalities, namely ECG, EDA, video, and audio. Nonetheless, sensors to collect ECG, EDA, video, and audio data are available for us at the Cyberpsychology Laboratory of UQO in the context of this project. The RECOLA data set was recorded at the University of Fribourg, Switzerland, to study socio-affective behaviors from multimodal data in the context of computer-supported collaborative work [7,18,24]. Spontaneous and naturalistic interactions were collected during the resolution of a collaborative task that was performed in pairs and remotely through a video conference. This consists of 27 5-minute synchronous audio, video, ECG and EDA recordings. Even though all participants speak French fluently, they have different nationalities (i.e., French, Italian or German), which provides some diversity in the expression of emotion. The data were labelled in the arousal and valence affective dimensions, and manually annotated using ANNEMO [24], an online, slider-based labelling tool. Each recording was annotated by six native French speakers. A combination of these individual ratings could be used as the ground truth label. The RECOLA data set was obtained, from [24], to 1) assist in the analysis of continuous

emotional dimensions, such as arousal and valence; and 2) for development and validation in the initial stage of the project.

2.3.3. State-of-the-Art Studies on RECOLA

This section presents relevant work from the literature on the prediction of arousal and valence values from physiological, visual, and multiple sensor sources, particularly focused on the RECOLA data set.

The authors of the original RECOLA data set [7] further extended their work in [23], where they performed experiments on the data set for the prediction of arousal and valence values. They extracted 54 physiological features from the ECG signal of RECOLA, and 60 physiological features form the EDA signal of RECOLA. They also extracted 20 visual features on each video frame in the video recordings of RECOLA along with their first order derivates (40 in total); as well as 65 acoustic features from the audio recordings of RECOLA along with their first order derivatives (130 in total). They then deployed a bidirectional long short-term memory (BiLSTM) RNN to predict arousal and valence measures. They compared the prediction performance of the RNN between mean ratings (average of annotations from all 6 raters) and all 6 ratings, using both single-task and multi-task learning techniques. For ECG, they achieved a CCC of 0.132 on arousal predictions, and a CCC of 0.149 on valence predictions using multi-task learning over the mean of the 6 ratings. For EDA, they obtained a CCC of 0.057 on arousal predictions, using single-task learning over the mean of the 6 ratings; and a CCC of 0.122 on valence predictions using multi-task learning over all 6 ratings. For visual data, they attained a CCC of 0.427 on arousal predictions, using multi-task learning over all 6 ratings. For valence, they reached a CCC of 0.431 using single-task learning over all 6 ratings. For acoustic data, they achieved a CCC of 0.788 on arousal predictions, using single-task learning over the mean of the 6 ratings. They obtained a CCC of 0.412 on valence predictions, using multi-task learning over the mean of the 6

ratings. The authors then performed various combinations of feature-level and decision-level fusion between the modalities. They attained a CCC of 0.804 on arousal predictions via decision-level fusion of the predictions obtained using the EDA, visual, and acoustic feature sets. For valence, they reached a CCC of 0.528 via decision-level fusion of the predictions obtained using the ECG, EDA, visual, and acoustic feature sets.

The same authors of [23] later introduced the Audio/Visual Emotion Challenge and Workshop (AVEC) in 2015 [25]. The baseline prediction performances achieved by the individual models for the gold-standard emotion sub-challenge (GES) in 2018's AVEC [26] are reported in terms of CCC. For EDA, the authors of [26] obtained a CCC of 0.029 and 0.058 for arousal and valence predictions, respectively. For ECG, they attained a CCC of 0.065 and 0.043 for arousal and valence predictions, respectively. These prediction performances were obtained through an emotion recognition system based on SVMs, used as static regressors. They also experimented with the different types of visual features: appearance, geometric, 17 facial AUs, and bags-of-words. For arousal, they reached a CCC of 0.312 via multi-task Lasso, while using appearance features. For valence, they achieved a CCC of 0.438 via an SVM, while using geometric features. Hierarchical fusion over the different data modalities was then applied and obtained a CCC of 0.657 for arousal and 0.515 for valence via Lasso and multi-task Lasso, respectively. The authors of [26] also experimented with acoustic data. More specifically, they utilized openSMILE tool to extract feature sets such as the Geneva minimalistic acoustic parameter set (GeMAPS), and its extended version (eGeMAPS), as well as the first 13 Mel-frequency cepstral coefficients (MFCCs) and their first and second derivatives. They then used the openXBOW tool to generate the XBoW representations of the acoustic features. Additionally, they used AlexNet to extract deep spectrum features from the speech files that they transformed into Mel-spectrogram images. Lastly, they tested multiple regression models on their

GeMAPS, deep spectrum, and bags-of-audio-words (BoAW) feature sets to predict arousal and valence values. As a result, they attained a CCC of 0.651 for arousal, and of 0.346 for valence predictions via an SVM trained on the BoAW feature set.

Amirian et al. [27] used random forests along with various schemes of fusion to predict arousal and valence values from RECOLA's acoustic, visual, and physiological features. For ECG, they reached a CCC of 0.097 and 0.139 on arousal and valence predictions, respectively. For EDA, they obtained a CCC of 0.074 and 0.206 on arousal and valence predictions, respectively. For visual features, they achieved a CCC of 0.516 using a forward echo state network (ESN) on arousal predictions, and 0.611 using a bidirectional ESN on valence predictions. In addition to RECOLA's acoustic features, the authors of [27] used the Log Frequency Power Coefficient (LFPC) with a log filter bank. They obtained a CCC of 0.759 using a bidirectional ESN on arousal predictions, and 0.335 using random forests on valence predictions. Their best prediction performances were obtained by a combination of random forests and linear regression fusion over all modalities (audio, visual, and physiological), namely a RMSE, PCC, and CCC of 0.118, 0.776, 0.762 for arousal, and 0.104, 0.634, 0.624 for valence; respectively.

The End2You tool [28] is a toolkit for multimodal profiling that was developed by the Imperial College of London to perform continuous dimensional emotion labels of arousal and valence values. It uses raw audio recordings, visual information (i.e., video), and physiological ECG signals as input. For ECG signals, it attained a CCC of 0.154 for arousal, and 0.052 for valence. For RECOLA's videos, it obtained a CCC of 0.358 for arousal, and 0.561 for valence. For RECOLA's audio recordings, it reached a CCC of 0.669 for arousal, and 0.286 for valence. The End2You tool achieved a CCC of 0.672 for arousal, and 0.521 for valence using multimodal data.

Brady et al. [29] used RECOLA's physiological data and baseline features, as specified in AVEC 2016 [18], to apply regression over arousal and valence values via a LSTM RNN. For ECG, they obtained an RMSE, PCC, and CCC of 0.218, 0.407, and 0.357 at predicting arousal values, respectively, and 0.117, 0.412, and 0.364 for valence values prediction, respectively. For EDA, they attained an RMSE, PCC, and CCC of 0.250, 0.089, and 0.082 at predicting arousal values, respectively. They reached an RMSE, PCC, and CCC of 0.124, 0.267, and 0.177 at predicting valence values, respectively. They also extracted higher level features from raw video and audio features using deep supervised and unsupervised learning, based on sparse coding, to improve the learning of the baseline SVM regressor. They used CNN features to predict arousal and valence values from video recordings, using an RNN. For video, they achieved an RMSE, PCC, and CCC of 0.201, 0.415, and 0.346; and 0.107, 0.549, and 0.511 on arousal and valence predictions, respectively. The authors of [29] also experimented with various types of acoustic features such as RECOLA's baseline features, MFCC features, shifted delta cepstrum (SDC) features, and prosody features. For arousal, they obtained an RMSE, PCC, and CCC of 0.107, 0.846, and 0.846 via a linear kernel SVM using the MFCC features. For valence, they attained an RMSE, PCC, and CCC of 0.132, 0.456, and 0.450 via a linear kernel SVM using the MFCC features. Finally, they proposed predicting continuous emotion dimensions using a state space approach such as Kalman filters, where measurements and noise are handled as Gaussian distribution. This was for fusing the emotional states (i.e., predictions) from the audio, video, and physiological data. After multimodal fusion, they reached a testing RMSE, PCC, and CCC of 0.115, 0.774, and 0.770; and 0.100, 0.689, and 0.687 on arousal and valence predictions, respectively. According to [24], the prediction performances obtained by the authors of [29] are the best prediction performances obtained on RECOLA in the literature. As such, the prediction performances in this thesis will be compared to theirs.

Han et al. [30] exploited the geometric visual features provided by AVEC to predict arousal and valence values through an RNN. They implemented an implicit fusion framework for joint audiovisual training. They achieved a CCC of 0.413 and 0.527 on arousal and valence predictions, respectively. Weber et al. [31] used visual features provided by RECOLA's team in 2016 to perform regression via a SVM with late subject, multimodal fusion (at decision/prediction-level). Their best CCCs were 0.682 and 0.468 for arousal and valence, respectively. Albadawy et al. [32] used the visual features provided by AVEC 2015, which included appearance (LGBP-TOP) and geometric (Euclidean distances between 49 facial landmarks) features. For arousal and valence predictions, they proposed a joint modelling strategy using a deep BiLSTM for ensemble and end-to-end models. Their ensemble BiLSTM model obtained a CCC of 0.699 and 0.617 for arousal and valence, respectively. As acoustic features, the authors of [32] used the Kaldi toolkit to extract a 40-dimensional log Mel filter bank coefficient. For arousal predictions from acoustic features, they attained a CCC of 0.697 using their end-to-end BiLSTM model. For valence predictions from acoustic features, they reached a CCC of 0.555 using their ensemble BiLSTM model.

The authors of [33] exploited CNN features from RECOLA's videos as well as an RNN to estimate valence values. They obtained an RMSE, PCC, and CCC of 0.107, 0.554, and 0.507; respectively. CNNs have also shown promising results when used to perform AR. AlexNet [8] was used in a number of studies to obtain deep visual features. AlexNet has also been applied on emotion recognition, where results demonstrated evident performance enhancements [9,10]. In this study, CNNs, such as ResNet and MobileNet, were exploited to predict continuous dimensional emotion annotations in terms of arousal and valence values from visual data.

Povolny et al. [34] presented a multimodal emotion detection algorithm using audio, bottleneck, and text-based features as well as the features suggested by Velstar et al.

[18]. The set of visual features in [18] were accompanied with CNN features, extracted from hidden layers, after training the CNN for landmark localization. The authors of [34] proposed multiple linear regression systems, trained on individual feature sets, for predicting the arousal and valence emotional dimensions. Their prediction performances achieved a CCC of 0.719 for arousal and 0.596 for valence. The CCC on the arousal results increased to 0.833 after the inclusion of bottleneck features on the validation data set of the RECOLA data set. In comparison to [34], Somandepalli et al. [35] used Kalman filters for decision level fusion. They first used support vector regression (SVR) to perform predictions from unimodal features, where predictions are noisy estimates of arousal and valence. The output of the SVR models was inputted to the Kalman filters for fusion. They later [36] proposed facial posture cues and a voicing probability scheme to deal with the multimodal nature of the problem. Their prediction performances obtained a CCC of 0.703 for arousal and 0.681 for valence. Table 1 summarizes the prediction performances of the above-mentioned state-of-the-art studies. A lower RMSE shows better performance. On the other hand, higher PCC and CCC indicate a better performance.

Table 1. Summary of Prediction Performances from the Literature on Predicting Arousal and Valence Values

Data Type	Prediction	Reference	Technique	RMSE, PCC, CCC
Physiological	Arousal	[23]	ECG Multi-task BiLSTM	N/A, N/A, 0.132
		[23]	EDA Single-task BiLSTM	N/A, N/A, 0.057
		[27]	ECG Random Forests	N/A, N/A, 0.097
		[27]	EDA Random Forests	N/A, N/A, 0.074
		[26]	ECG SVM	N/A, N/A, 0.065
		[26]	EDA SVM	N/A, N/A, 0.029
		[28]	ECG End2You	N/A, N/A, 0.154
		[29]	ECG RNN	0.218, 0.407, 0.357
		[29]	EDA RNN	0.250, 0.089, 0.082
	Valence	[23]	ECG Multi-task BiLSTM	N/A, N/A, 0.149
		[23]	EDA Multi-task BiLSTM	N/A, N/A, 0.122
		[27]	ECG Random Forests	N/A, N/A, 0.139
		[27]	EDA Random Forests	N/A, N/A, 0.206
		[26]	ECG SVM	N/A, N/A, 0.043
		[26]	EDA SVM	N/A, N/A, 0.058

Data Type	Prediction	Reference	Technique	RMSE, PCC, CCC
		[28]	ECG End2You	N/A, N/A, 0.052
		[29]	ECG RNN	0.117, 0.412, 0.364
		[29]	EDA RNN	0.124, 0.267, 0.177
Visual	Arousal	[23]	Multi-task BiLSTM	N/A, N/A, 0.427
		[27]	Forward ESN	N/A, N/A, 0.516
		[26]	Multi-task Lasso	N/A, N/A, 0.312
		[28]	End2You	N/A, N/A, 0.358
		[29]	CNN + RNN	0.201, 0.415, 0.346
		[30]	RNN	N/A, N/A, 0.413
		[31]	SVM + Subject Fusion	N/A, N/A, 0.682
		[32]	Ensemble BiLSTM	N/A, N/A, 0.699
	Valence	[23]	Single-task BiLSTM	N/A, N/A, 0.431
		[27]	Bidirectional ESN	N/A, N/A, 0.611
		[26]	SVM	N/A, N/A, 0.438
		[28]	End2You	N/A, N/A, 0.561
		[29]	CNN + RNN	0.107, 0.549, 0.511
		[30]	RNN	N/A, N/A, 0.527
		[31]	SVM + Subject Fusion	N/A, N/A, 0.468
		[32]	Ensemble BiLSTM	N/A, N/A, 0.617
		[33]	CNN + RNN	0.107, 0.554, 0.507
Acoustic	Arousal	[23]	Single-task BiLSTM	N/A, N/A, 0.788
		[26]	SVM	N/A, N/A, 0.651
		[27]	Bidirectional ESN	N/A, N/A, 0.759
		[28]	End2You	N/A, N/A, 0.669
		[29]	Linear SVM	0.107, 0.846, 0.846
		[32]	End-to-End BiLSTM	N/A, N/A, 0.697
	Valence	[23]	Multi-task BiLSTM	N/A, N/A, 0.412
		[26]	SVM	N/A, N/A, 0.346
		[27]	Random Forests	N/A, N/A, 0.335
		[28]	End2You	N/A, N/A, 0.286
		[29]	Linear SVM	0.132, 0.456, 0.450
		[32]	Ensemble BiLSTM	N/A, N/A, 0.555
Multimodal	Arousal	[23]	Decision-level Fusion (BiLSTM)	N/A, N/A, 0.804
		[27]	Random Forests + Linear Regression	0.118, 0.776, 0.762
		[26]	Hierarchical Fusion + Lasso	N/A, N/A, 0.657
		[28]	End2You	N/A, N/A, 0.672
		[29]	Kalman Filters	0.115, 0.774, 0.770
		[34]	Multiple Linear Regressors	N/A, N/A, 0.833
		[36]	SVR + Kalman Filters	N/A, N/A, 0.703
	Valence	[23]	Decision-level Fusion (BiLSTM)	N/A, N/A, 0.528
		[27]	Random Forests + Linear Regression	0.104, 0.634, 0.624
		[26]	Hierarchical Fusion + Multi-task Lasso	N/A, N/A, 0.515
		[28]	End2You	N/A, N/A, 0.521
		[29]	Kalman Filters	0.100, 0.689, 0.687

Data Type	Prediction	Reference	Technique	RMSE, PCC, CCC
		[34]	Multiple Linear Regressors	N/A, N/A, 0.596
		[36]	SVR + Kalman Filters	N/A, N/A, 0.681

The following references showed the potential of using RNNs and CNNs to perform AR. Gunes and Schuller [19] trained two separate deep CNNs. The CNNs were pre-trained on a large data set, and then fine-tuned on a data set of audio and video. Gunes et al. [37] showed the potential of using LSTMs for dimensional emotion prediction. Ringeval et al. [38] used a LSTM RNN to perform regression for dimensional emotional recognition based on visual, audio, and physiological modalities. Chen et al. [39] used LSTMs to identify the long-term inter-dependency within segments of a multimedia signal. They proposed a new conditional attention fusion scheme, in which modalities are weighted according to their current and previous features. Tzirakis et al. [40] used a shallow network followed by identity mapping to extract features from raw audio and video signals. The obtained features were then inputted into a two-layer LSTM. The LSTM was trained from end-to-end instead of training it on individual components separately. This approach outperformed traditional approaches based on baseline handcrafted features in the RECOLA data set. Huang et al. [41] applied a deep neural network and hypergraphs for emotion recognition using facial features. Facial features were extracted from the last fully connected layer of the trained CNN, which were then treated as attributes for the hypergraph. Kahou et al. [42] used a CNN to extract features that were input into an RNN. The RNN categorizes the emotions in RECOLA's video recordings.

This literature review shows the potential of using classical machine learning (e.g., SVM and linear regression) or deep machine learning regression (e.g., CNNs and RNNs) in AR solutions. As can be seen in Table 1 above, most state-of-the-art studies performed complex processing, feature extraction, and multimodal fusion processes. Despite these efforts, the prediction performances of their models can still be improved.

In this work, simpler processing techniques were combined with novel feature selection and decision fusion mechanisms, and better prediction performances were attained using similar recordings of RECOLA, as will be discussed and demonstrated in the remainder of this thesis.

2.4. Virtual Reality (VR)

VR is a technology that creates a simulated environment that mimics the real world, enabling users to interact with it in a lifelike manner [43]. It employs both hardware and software to generate a 3D space that can be explored in 360 degrees. Users typically wear a VR HMD or a similar device to engage with this environment effectively, isolating them from their physical surroundings. The HMD delivers visuals to each eye, creating a sense of depth, while head and body tracking synchronize the movements of the user with the virtual world. VR can be used for various applications, including entertainment (e.g., video games), education (e.g., medical or military training), and business (e.g., virtual meetings). By providing a sense of presence in a virtual space, VR can make these experiences more engaging and realistic.

In other words, VR serves as a visualization tool that minimizes external distractions. VR systems typically include HMDs that cover the user's eyes, helping to further reduce distractions [5]. VR restrains visual feedback and applies interactivity methods to create a sense of presence, making users feel like they are in the presented environment. The immersion of senses benefits many fields of study, which can range from entertainment and education to psychological therapy. Interactivity in VR creates a connection between the application and the user. VR interprets movements from the physical world to the virtual world in a precise manner. This can be designed via controls that simulate physical actions, allowing the user to attain positional awareness easily. One can benefit from this paradigm by virtually enabling visual representations

of workflows and data. VR implements intuitive movements for interaction, in an environment which has the ability to represent 3D objects. VR provides a method of visualization that has the potential for mimicking day-to-day life.

In the field of mental health, VR can address some of the limitations of traditional cognitive remediation interventions [43]. Cognitive remediation refers to the use of techniques to teach and practice cognitive skills (i.e., better "thinking skills"). By providing an immersive experience that simulates 3D environments, VR enables the creation of cognitive remediation treatments with higher ecological validity. Research on the acquisition of skills other than cognitive ones indicates that immersive VR programs are more effective than non-immersive ones in facilitating the transfer of learned skills to real-world contexts. These findings emphasize the potential therapeutic benefits of immersive VR technology for mental health treatments.

Preliminary evidence suggests that VR-based cognitive remediation can enhance positive symptoms, attention, memory, spatial learning, and social skills in people with schizophrenia [43]. Additionally, people with schizophrenia find VR-based cognitive remediation enjoyable and engaging, making it a potentially more effective method for delivering cognitive remediation.

VR could serve as an innovative method to enhance the real-world applicability and transfer of learned skills acquired through traditional cognitive remediation programs [43]. This thesis sought to evaluate the feasibility of a VR-based cognitive remediation module to improve the application of AR in people with schizophrenia.

2.5. Affect Recognition (AR) in Virtual Reality (VR)

VR systems and characters can be exploited to analyze the facial emotions of people with mental disorders such as schizophrenia and autism [44]. Hallucinations, delusions, thought disorders, and abnormal social and affective behaviours are all symptoms of schizophrenia, which influence cognitive, emotional, and social functions [44-46]. The affective state of people with schizophrenia while perceiving emotions are often overlooked. Most state-of-the-art rely on the self-reporting of emotional experiences in watching videos [47,48] and images [49] to perform AR. In real life, we express, perceive, and experience emotions dynamically over time. VR interfaces can simulate the temporal nature of emotional expressions. VR technology has potential in schizophrenia intervention as it offers flexibility, controllability, and adaptability.

Utilizing VR technology alongside physiological and eye gaze monitoring provides more insight on emotion processing than static images and videos [44]. Bekele et al. [44] presented a VR system which controls the facial expressions of a virtual character with an eye-tracker and physiological sensors such as PPG, skin temperature (SKT), galvanic skin response (GSR), EMG, and respiration (RSP) to better understand how the patients of schizophrenia recognize emotions. 81 features were extracted from PPG, GSR, RSP, SKT, and EMG for this analysis. According to the psychophysiology literature, the selected features correlate with engagement and emotion recognition processes [50-53]. The heart rate (HR) was extracted from the PPG signal. The GSR was partitioned into its phasic and tonic components, from which the skin conductance response (SCR) rate and mean skin conductance level (SCL) were extracted. The breathing rate (BR) was extracted from the RSP signal. The mean skin temperature was extracted from the SKT signal. The EMG signal was used for the recognition of facial emotions. The pupil diameter, fixation duration, sum of fixation counts (SFC), saccade path length (SPL), and blink rate were extracted from the eye tracking data.

The two clustering methods: k-means and density peaks were used to analyze physiological data. The cluster analysis showed that GSR (a.k.a., EDA) was the best predictor and that it was more associated with arousal than valence.

The Affective VR System (AVRS) [54] contains numerous emotionally evocative VR scenes, which are rated in terms of their affectivity. AVRS better stimulates affects of happiness, sadness, fear, relaxation, disgust, and rage emotions, as it omits environmental interventions. The affective VR scenes of AVRS have been created from affective pictures, videos, and audios. The three affective dimensions of valence, arousal, and dominance/control were considered in the emotional assessments, where the Self-Assessment Manikin (SAM) [55] rating system was utilized. 100 participants used SAM to rate the 8 emotion-evoking VR scenes of the AVRS based on the dimensions of valence, arousal, and dominance.

Joyal et al. [56] used a set of virtual faces expressing happiness, surprise, anger, sadness, fear, and disgust emotions. The virtual faces were tested and compared against real faces for emotion recognition. The data from facial EMGs (zygomatic major and corrugator supercilii muscles), and regional gaze fixation latencies (eyes and mouth regions) were also collected and compared. They created 24 videos: 2 (real and virtual) × 2 (man and woman) × 6 (emotions). The videos consisted of real and synthetic expressions, presented in 10 seconds, led by 2 seconds for central cross fixation. Overall, the recognition rates did not significantly differ between real and virtual expressions, nor for each emotion. No difference emerged between the mean contractions of the zygomatic major or the corrugator supercilii muscles between both real and virtual conditions for any emotions. Finally, only the time spent looking at the mouth differed significantly between real and virtual conditions.

In [57], human and virtual faces were assessed, by 32 healthy volunteers, for their expressed emotions. The faces expressed emotions such as happiness, sadness, anger, fear, disgust, and neutral. The obtained results show that both human and virtual faces can be recognized equally. The results are dependent on the emotion. Disgust is generally harder to convey virtually. On the other hand, virtual faces achieved better recognition for the emotions of sadness and fear than human faces. The age of volunteers also played a role in recognition rates.

Benlamine et al. [58] introduce a framework designed to adapt game elements based on the player's emotional state, integrating this framework into a VR environment aimed at moral development. The game elements considered include avatar gestures and facial expressions during player interactions, background music, scoring, game mechanics, aesthetics, and learning components. The framework, named BARGAIN (Behavioral Affective Rule-based Games Adaptation Interface), served as an authoring tool for affective game design. It featured a visual interface that uses the finite state machine (FSM) technique to depict affective rules as state transitions, which depend on the player's emotional state assessed through a facial expression recognition system based on EEG data. A user study involving 29 participants was conducted to evaluate the impact of the affective VR game on player experience using a game experience questionnaire (GEQ). The study found a significant correlation between GEQ dimensions and players' facial expressions during interactions with non-player characters (NPCs) in the VR game, indicating that adapting games to users' emotions enhances their experience.

The authors of [59] conducted experiments to predict players' emotions during video game sessions using physiological data. They collected data from 33 participants playing games on a standard monitor and a VR HMD, creating a data set that included ECG, five facial EMGs, EDA, and RSP. Participants also self-assessed their emotional

state in terms of arousal and valence. The researchers analyzed the contribution of each physiological signal to the players' mental state and applied machine learning algorithms to predict emotional states. The results, showing low errors in predictions, validated the experimental framework. The used algorithm can accurately predict self-assessed emotions in the arousal/valence 2D space, potentially aiding in the development of game devices and engines that detect physiological data and enhance game design.

In [60], a VR experiment driven by narrative content utilized novel biosensing technology to assess emotional responses to a complex soundscape, which included discrete and ambient sounds, music, and speech. The stimuli were presented in both spatialized and mono audio formats to determine if head-tracked spatial audio influences physiologically measured emotional responses. The study also examined how the spatial characteristics of audio affect a listener's sense of immersion in a VR environment by analyzing physical movement data. The findings indicate that spatial audio significantly impacts emotional responses in immersive virtual environments (IVEs). Additionally, physical movement data revealed increased user intention and focused localization with spatial audio compared to non-spatial audio.

Most state-of-the-art studies concentrate on comparing the emotions of virtual and human faces, or analyze the correlation between predictors and arousal/valence. Some studies from the literature assessed the effects of certain settings on the users. They showed that adjusting the settings of virtual environments based on the affective state of the user can really enhance VR experiences. This Ph.D. thesis focused on predicting the emotional affective states of participants during VR immersions through the arousal and valence measures. These predictions can be used in the future to control VR environments accordingly.

3. PROPOSED SOLUTION

The goal of this thesis is to design and develop a novel adaptable intervention to remediate cognitive impairments in people with schizophrenia using VR, based on synergistic computer science and psychology approaches. A novel machine learning approach which depends on visual and physiological sensory data is required to adaptively adjust the virtual environment to the affective states of users. In the future, the aim is to also automatically optimize the level of cognitive effort requested by users, while avoiding discouragement. In the proposed solution, a multi-sensory system was used to improve the current state of the art solutions for the prediction of affective states. Information from the various data sources in the system were treated to predict the affective state of the user, as if they were immersed in VR through supervised, classical and deep machine learning techniques. Figure 2 displays a high-level diagram of the proposed solution. It is composed of subsystems: one system for each data modality, namely the visual (video), acoustic (audio), and physiological (EDA and ECG) data modalities. The focus was on one data modality at a time to perfect the results for each modality first, before combining all modalities in the final system (i.e., multimodal fusion). Data from RECOLA were initially used as a proof-of-concept mechanism to propose solutions that can be compared with the existing literature. Finally, real data that were collected at the Cyberpsychology Laboratory of UQO were operated on.

As further presented in the following sections, the physiological signals, acoustic features, and visual recordings of RECOLA were processed and labelled. Then, continuous dimensional emotion predictions of arousal and valence values were performed, through classical machine learning and deep learning techniques.

Specifically, the experiments included tree and ensemble regression for physiological and acoustic data. Additionally, a BiLSTM RNN was tested on physiological data. Tree and ensemble models were chosen because they are known to be fast and effective [13]. On the other hand, a BiLSTM RNN was chosen because it is capable of processing data in the form of time series and sequences, which suits the needs of this study. BiLSTM RNNs can learn from past and future samples/time steps. For visual data, experiments involved deep machine learning methods such as ResNet-18 and MobileNet-v2 CNNs as they offer a good trade-off between accuracy, speed, and size. Experiments using classical machine learning methods such as tree and ensemble regression were also performed on visual data.

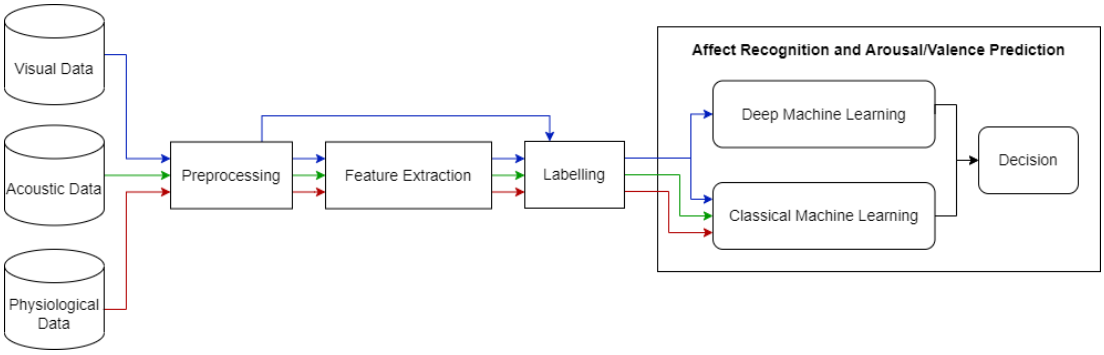


Figure 2. An overview of the proposed solution. The red line denotes physiological data, the blue line denotes visual data, and the green line denotes acoustic data.

3.1. Overview

The RECOLA data set includes recordings of 27 participants, from which the recordings of 18 participants contain all types of data modalities (i.e., audio, video, ECG, and EDA). These 18 were used for training and validation, as well as testing purposes. The data were not separated per participant/person throughout this study. All data were combined and shuffled to ensure data diversity, in terms of emotions, during the training process. Most of works in the literature proceeded in the same manner. An initial approach was attempted where the data per participant/person were separated,

but that led to poor results. Appendix A presents some of these results. The fact that RECOLA contains data from few participants has made it hard to generalize the data as there was no sufficient diversity of people. Also, arousal and valence measures could differ significantly from one person to another. Hundreds of participants might be required in order to generalize the data per participant better.

First, physiological data from RECOLA's physiological signal recordings of EDA and ECG were operated on. The EDA and ECG signals were processed by applying time delay, early features fusion, feature processing, arousal and valence annotation labelling, and data shuffling and splitting. Early fusion was used to combine the EDA and ECG modalities at the feature level. Tree and ensemble regressors were exploited for the purpose of predicting continuous dimensional emotion annotations in terms of arousal and valence values. Afterwards, the work on physiological data was further extended by (1) adding preprocessing operations; (2) applying feature standardization; (3) applying feature selection; (4) exploring RNNs, specifically BiLSTM; and (5) introducing decision fusion.

Secondly, visual data from RECOLA's video signal recordings were operated on. Videos were preprocessed by applying frame extraction and sequencing, face detection and cropping, annotation labelling, and data augmentation. After processing, the extracted images (i.e., video frames) of participants' faces were inputted into CNNs, such as ResNet-18 and MobileNet-v2, for predicting arousal and valence values.

Thirdly, acoustic features from RECOLA were extracted. The acoustic features were processed by applying time delay, feature processing and selection, arousal and valence annotation labelling, and data shuffling and splitting. Tree and ensemble regressors were utilized to predict continuous dimensional emotion annotations in terms of arousal and valence values.

Lastly, multimodal AR was performed by combining multimodal (physiological, acoustic, and visual) features, and/or combining arousal/valence predictions through a decision fusion mechanism. After combining multimodal features, machine learning models such as tree and ensemble regressors were trained to predict arousal and valence annotations.

After training machine learning models, they were tested by predicting the arousal and valence values on the testing sets. This helped to evaluate the performance of the trained models when they were presented with new data. The RMSE, PCC, and CCC performance measures were used to enable a comparison with similar work in the literature.

3.2. Leveraged Machine Learning Models

As aforementioned, tree and ensemble regressors, as well as a BiLSTM RNN were used to perform AR on the physiological data recordings of the RECOLA data set as a proof-of-concept. Various tree and ensemble regression models were experimented with. More specifically, fine, medium, coarse, and optimizable regression trees, as well as boosted trees, bagged trees, and optimizable ensemble regression models were used to predict the arousal and valence values from EDA and ECG samples. A BiLSTM RNN was used to predict the arousal and valence values from EDA and ECG sequences.

On the other hand, tree, kernel, and ensemble regressors, as well as pretrained CNNs were used to perform AR on the visual data recordings of the RECOLA data set. More specifically, fine, medium, coarse, and optimizable regression trees; SVM and least squares regression kernels; as well as the boosted trees, bagged trees, and optimizable ensemble regression models were used to predict the arousal and valence values form predesigned visual features. The ResNet-18 and MobileNet-v2 CNNs were used to

predict arousal and valence values from face image. The following sections will briefly explain the architectures of these models, and how they operate.

For acoustic features, tree and ensemble regressors were used to perform AR on the acoustic features from the RECOLA data set as a proof-of-concept. Fine, medium, coarse, and optimizable regression trees, as well as boosted trees, bagged trees, and optimizable ensemble regression models were used to predict the arousal and valence values from acoustic features.

3.2.1. Regression Trees

Regression trees are usually binary, where each step in the prediction processes contains a check condition on the value of one attribute/feature of the input data [13]. They can be trained to predict the responses to the input data. To predict the response of a regression tree, one can start from the root (beginning) node of the tree and go down until a leaf node is reached. At each node, a decision, based on the rule(s) of the check condition associated with that node, is taken to find the next branch. This pattern continues until one arrives at a leaf node. The predicted response would then be the value at the leaf node. Figure 3 shows an example of a regression tree. This tree predicts the response based on the values of two attributes/features, $x_1$and x_2.

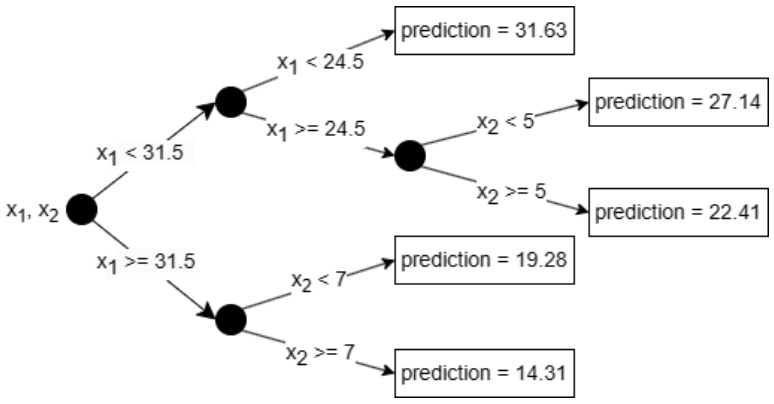


Figure 3. Example of a regression tree [13].

The Regression Learner app in MATLAB uses the fitrtree function to train regression trees [13]. This function provides options such as minimum leaf size and surrogate decision splits. The minimum leaf size specifies the minimum amount of training samples needed to predict the responses of leaf nodes. A fine regression tree is small and has a minimum leaf size of 4. A medium regression tree has a minimum leaf size of 12. A coarse regression tree is larger, with a minimum leaf size of 36. An optimizable regression tree optimizes training hyperparameters (i.e., minimum leaf size) using a Bayesian optimizer. A fine tree with many small leaves tends to be very accurate on the training data, but it might show a lower performance on an independent test set. A tree with many leaves is more prone to overfitting, and its validation performance is usually much lower than its training performance. Unlikely, a coarse tree with less large leaves would not achieve a high training performance. However, its testing performance can be comparable to its training performance.

The surrogate decision splits option specifies surrogate use for decision splits [13]. It is used when there are missing data to improve the accuracy of predictions. When the surrogate decision splits option is on, the regression tree finds surrogate splits at each branch node. For the purposes of this work, the surrogate decision splits option was kept turned off. Regression trees were chosen for their easy interpretation, fast speed at fitting and prediction, and low memory usage [13].

3.2.2. Kernel Regression Models

Kernel regression models perform nonlinear regression on data with many samples [13]. Regression kernels are Gaussian regression models for nonlinear regression over large data sets. Gaussian kernel regression models map attributes/features in a low-dimensional space into a high-dimensional space, and then fit a linear model to the transformed attributes/features in the expanded, high-dimensional space. There are two fitting options for kernel regression: an SVM linear model and a least-squares linear

model. An SVM kernel maps the attributes/features into a high-dimensional space and fits a linear SVM model to the transformed attributes/features. A least squares regression kernel maps the attributes/features into a high-dimensional space and fits a least squares linear regression model to the transformed attributes/features. Kernel regression models were chosen because they tend to train and predict faster for large data sets than other models such as SVMs [13].

The fitrkernel function was used to train kernel regression models [13]. This function provides options such as learner, number of expansion dimensions, regularization strength (lambda (λ)), kernel scale, epsilon (ε), and iteration limit. The learner option specifies whether to fit the data into an SVM or least squares linear regression model. An SVM kernel uses an epsilon-insensitive loss, while the least-square kernel uses a mean squared error (MSE) during model fitting. The number of expansion dimensions specifies the dimensions of the expanded, high-dimensional space. The number of dimensions is automatically set to $2.\text{^}ceil(min(\log_2(p) + 5, 15))$, where p is the number of attributes/features. The regularization strength parameter, λ, specifies the ridge $(L2)$ regularization penalty term. When an SVM learner is selected, the box constraint, C, and the regularization term strength λ are correlated by $C = 1/(\lambda n)$, where n is the number of samples. The box constraint, C, controls the penalty imposed on samples with large residuals. A greater value of C provides a more flexible model. A smaller value of C provides a less flexible model, which is less prone to overfitting.

The regularization strength was automatically set to $1/n$, where n is the number of samples. The kernel scale option specifies the kernel scaling, which is used to configure the random feature expansion. A heuristic procedure, based on subsampling was used to select the kernel scale value. Epsilon (ε) specifies half the width of the epsilon-insensitive band of the SVM kernel. The value of epsilon was defined as $iqr(Y)/13.49$, an estimate of a tenth of the STD using the interquartile range of the prediction variable,

Y. In the case of this study, Y is represented by either the arousal or valence value. If $iqr(Y)$ is equal to zero, the value of epsilon was set to 0.1. Finally, the iteration limit parameter specifies the maximum number of training iterations. Note that all of these options/parameters can either be set automatically as described above, or manually by specifying a value.

3.2.3. Ensemble Regression Models

Ensembles of regression trees combine results from multiple learners into one ensemble model [13]. There are three ensemble learning algorithms: bagging, random space, and various boosting algorithms. Three ensemble models were utilized: bagged trees, boosted trees, and optimizable ensemble. Bagged trees represent a bootstrap-aggregated ensemble of regression trees. Boosted trees represent an ensemble of regression trees using the least-squares boosting (LSBoost) algorithm. An optimizable regression ensemble optimizes training hyperparameters (ensemble method, number of learners, learning rate, minimum leaf size, and number of predictors to sample) using Bayesian optimization.

Bagging is an ensemble method that decreases the influence of overfitting and improves generalization via bootstrap aggregation [13]. The bootstrap aggregation (bagging) method bags a weak learner (e.g., decision tree) on a data set, and generates various bootstrap copies of the data set. Then, it grows decision trees on the copied data sets. The bootstrap copies are obtained by randomly selecting N out of N samples with replacement, where N is the size of the input data set. Every tree in the ensemble model can randomly select features/attributes for each decision split, this technique is known as random forest [13,61]. The random forests technique has been proven to enhance the performance of bagged trees. The number of features/attributes to be randomly selected for each split was set to one third of the number of features/attributes. Therefore, the bagged trees model grows every tree using bootstrap samples of the

input data. Readings not included in a bootstrap sample are considered “out-of-bag” for that tree. The model selects a random subset of features/attributes for each decision split by using the random forest algorithm [61].

The fitrensemble function was used to train ensemble models [13]. This function provides options such as minimum leaf size, number of learners, learning rate, and the number of features/attributes to sample. The minimum leaf size specifies the amount of training samples used to predict the response at leaf nodes. Changing the number of learners can improve the performance of the model. Many learners can achieve higher performances, but their training can be time consuming. The number of predictors to sample specifies how many features/attributes are selected at random for each branch in the learning trees. All available features/attributes were selected.

The minimum number of samples per leaf for bagged trees was set to 5 [13]. Trees that are grown using the default leaf size are often very deep. These settings are almost optimal for the predictive power of ensembles. Usually, growing trees with larger leaves can be trained without losing predictive power; and thus, reduce training and prediction time, and memory usage for the trained ensemble. The LSBoost ensemble method minimizes the MSE by fitting a new learner at every step of the training process. The new learners are fit to the difference between the observed response and the aggregated prediction of all previous learners. The LSBoost method can be used with shrinkage by passing in the learning rate parameter. The default value of this parameter is set to 1, where the ensemble learns at maximum speed. Alternatively, one can optimize the abovementioned hyperparameter using an optimizable ensemble.

3.2.4. BiLSTM RNN

An LSTM neural network is a type of RNNs which learns long-term dependencies between time steps of sequence data [13]. An LSTM network processes input data by

iterating through time steps and updating the state of the network. The network state holds information remembered from the previous time steps. A BiLSTM RNN learns bidirectional long-term dependencies between time steps of sequence data [13]. This is useful for learning from the complete sequence at each time step. The main layers of an LSTM neural network include a sequence input layer and an LSTM layer. A sequence input layer inputs sequence data into the neural network. An LSTM layer learns the long-term dependencies between the time steps of input sequences. Figure 4 illustrates the architecture of a simple BiLSTM neural network for regression.

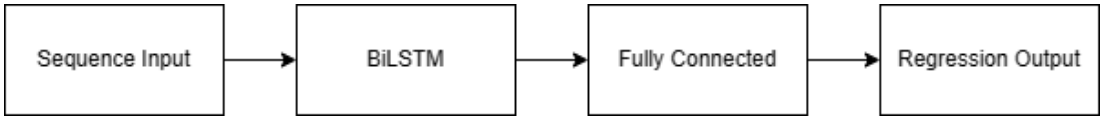


Figure 4. Architecture of a simple BiLSTM network [13].

The first layer of the neural network is a sequence input layer, while the second layer is a BiLSTM layer [13]. To predict response values, the neural network ends with a fully connected layer and a regression output layer. A fully connected layer multiplies the input data by a weight matrix; then, adds a bias vector to it. All neurons in a fully connected layer connect to all neurons in the previous, BiLSTM layer. The fully connected layer combines all of the information learned by previous layers to find patterns in the data. Fully connected layers reshape the output to encode the spatial data in the channel dimension. The regression output layer calculates the half-mean-squared-error loss of regression tasks.

3.2.5. Pre-Trained CNNs

For a number of applications, using a simple network with a sequence of layers is adequate, while a more complex network might be needed for other applications [13]. A more complex network would consist of layers that accept inputs from multiple layers and send outputs to multiple layers. Such networks are called directed acyclic

graph (DAG) networks. He et al. introduced residual networks (ResNets) in 2015 [62] for image classification. A ResNet is a DAG network containing residual, shortcut connections which bypass the layers of the main network. In other words, “these connections allow the input to skip the convolutional units of the main branch; thus, providing a simpler path through the network” [13]. Such connections allow the gradients of parameters to easily flow from the output layer up to the prior layers of the network. This enables the training of deeper networks as it resolves the problem of vanishing gradients during the early stages of training. Moreover, an increased network depth can result in better prediction performances for complicated tasks.

The structure of a ResNet is composed of initial layers, followed by stacks containing residual blocks, then the final layers [13,62]. There are three kinds of residual blocks: the initial, standard, and downsampling. Each residual block encompasses deep learning layers. The layers of each block depend on the type of the block and the options set by the developer. The initial block is located at the beginning, before the first stack. The layers of this block determine whether the block holds activation sizes or applies downsampling. The standard residual block follows the downsampling residual block in each stack. This block occurs multiple times in each stack and maintains the activation sizes. The downsampling residual block occurs once at the start of each stack, subsequent to the first stack. The first convolutional unit in this block downsamples the spatial dimensions by a factor of 2. The depth of each stack may differ. Figure 5 shows an example of a ResNet that has two stacks of decreasing depth. The first stack has a depth of 3, whereas the second stack has a depth of 2.

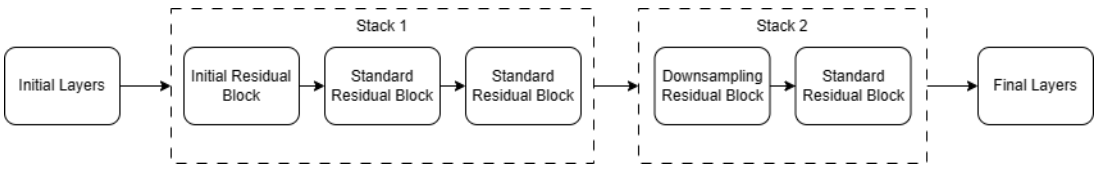


Figure 5. Example of a ResNet structure [13].

The initial, standard, and downsampling residual blocks can either be bottleneck or non-bottleneck blocks [13,62]. Bottleneck blocks run a 1 × 1 convolution, before running a 3 × 3 convolution, to lower the number of channels by a factor of 4. Networks with and without bottleneck blocks have similar computational complexities. However, networks with bottlenecks are 4 times larger in terms of the total number of features/attributes travelling in the residual connections. Bottleneck blocks help to boost the efficiency of networks.

ResNet-18 is one example of ResNets. ResNet-18 is a 72-layer architecture with 18 deep layers [62]. The architecture of this network enables large amounts of convolutional layers to function efficiently. MobileNet-v2 is an inverted ResNet that uses linear bottlenecks [63]. It is a 154-layer architecture with 53 deep layers. The inverted residual structure of MobileNet-v2 has residual, shortcut connections between bottleneck layers. The input and output layers of residual blocks in MobileNet-v2 are thin bottleneck layers, unlike the original ResNet structure, where expanded layers are inputted. A MobileNet-v2 uses lightweight, depth-wise convolutions to filter features/attributes in the intermediate expansion layers. MATLAB provides pretrained ResNet-18 and MobileNet-v2 CNNs.

3.3. Unimodal and Multimodal Decision Fusion

The testing predictions from various regressors and neural networks were fused by averaging them to observe how this fusion affected the prediction performance. Let N be the number of trained regressors and neural networks, and P be the predictions set obtained by model i; the final predictions set, P_{final}, can then be computed as follows:

$$P_{final} = \frac{P_1 + P_2 + \cdots + P_n}{N} = \frac{\sum_{i=1}^{n} P_i}{N} \quad (4)$$

This method provides a unified process to fuse the predictions of multiple regression models at the decision level. Nonetheless, this method is effective on all data modalities: physiological, visual, and acoustic.

4. PROOF-OF-CONCEPT FOR AFFECT RECOGNITION USING PHYSIOLOGICAL DATA

This chapter presents stages 2 to 5, and 13 of this research and is related to contributions 1, 4, and 5 from Section 1.3. It describes the proposed solution for the use of physiological data in continuous emotional predictions of arousal and valence values. As a reminder, RECOLA's EDA and ECG recordings were operated on as physiological data.

As mentioned before, the RECOLA recordings of 18 participants were used for training and validation, as well as testing purposes. All records were preprocessed by applying time delay and sequencing, early feature fusion, replacement of missing values, feature standardization/normalization, feature scaling, feature selection, arousal and valence annotation labelling, and data shuffling and splitting, as it will be further presented in the next subsections. After processing the physiological data, classical regressors and a BiLSTM RNN were used for predicting arousal and valence values. Figure 6 details the proposed approach for physiological data.

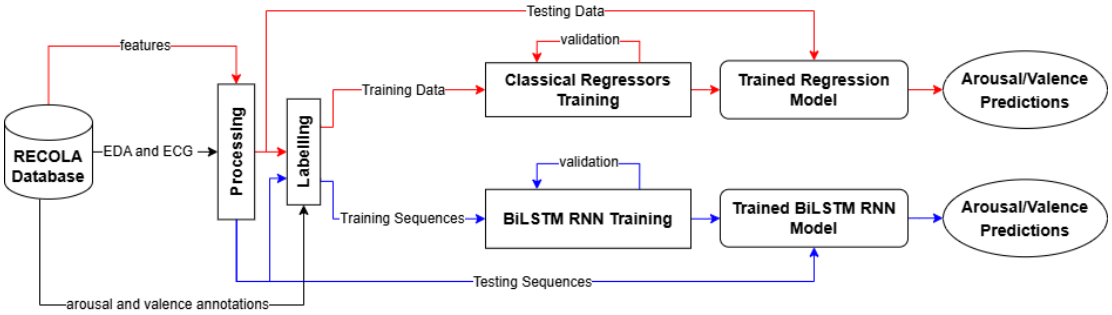


Figure 6. An overview of the proposed approach for physiological data. For physiological data, classical regressors as well as a BiLSTM RNN were experimented with to predict arousal and valence values. The red lines represent the flow for the classical regression approach, and the blue

lines represent the flow for the BiLSTM RNN approach. The black lines represent mutual steps between the two approaches. Note that the predictions from the two approaches are separate and are not combined.

4.1. Preprocessing of Physiological Data

This section details the processing steps that were applied to the physiological recordings of RECOLA. Figure 7 shows a breakdown of these processing steps that are contained in the "Processing" block of Figure 6. This work corresponds to stages 2 and 5 of this research and is related to contributions 1 and 4 (see Section 1.3).

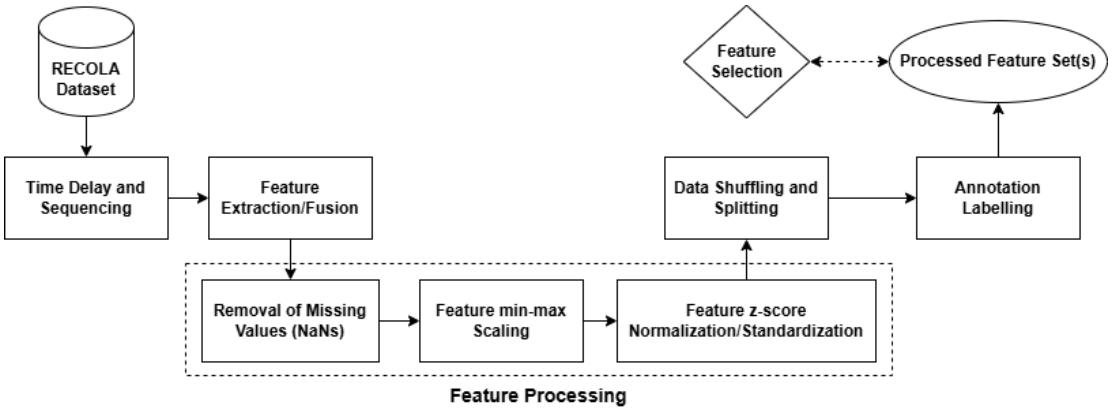


Figure 7. Breakdown of processing steps for physiological data.

4.1.1. Time Delay and Sequencing

RECOLA's physiological recordings (i.e., EDA and ECG) were sampled at a rate of 1000 samples per second. This means that one sample was captured every 1 millisecond. On the other hand, RECOLA's audio and video recordings were sampled every 40 milliseconds. Similarly, the physiological features were calculated every 40 milliseconds as well. To enable the synchronous use of data, the EDA and ECG signals were subsampled by considering only the readings occurring every 40 milliseconds. However, the corresponding EDA and ECG features in RECOLA were only calculated after 2 seconds of recording. Therefore, the readings that were collected before that time were skipped. As a result, the first 50 samples of the recordings were discarded.

4.1.2. Early Feature Fusion

The baseline EDA and ECG features of RECOLA, as described in [18], were used. SCR is EDA's rapid, transient response, whereas SCL is EDA's slower, basal drift. In [18], both SCR and SCL were extracted from the EDA signal through a third-order Butterworth filter, at different cut-off frequencies. In addition to the EDA readings, there are 62 EDA features, which include:

- the slope;
- the fast Fourier transform (FFT) entropy and mean frequency of the SCR;
- the mean, its first-order derivative, and the negative part of its derivative for EDA, SCR, and SCL;
- the STD, kurtosis, and skewness of EDA, SCR, and SCL;
- the proportion of EDA, SCR, and SCL;
- the x-bound of EDA, SCR, and SCL;
- the non-stationary index (NSI) and normalized length density (NLD) of EDA, SCR, and SCL; and
- deltas of all of the above.

Besides the ECG readings, there are 54 ECG features, which consist of:

- the HR and HR variability (HRV);
- the zero-crossing rate;
- the first 12 FFTs;
- the entropy, mean frequency, and slope of ECG's FFT;
- the first four statistical moments (mean, STD, kurtosis, and skewness);
- NSI and NLD;
- the power at very low, low, high, and low/high frequencies; and
- deltas of all of the above.

These features were provided in the format of .arff files. The .arff files were converted into .csv files to easily load them into MATLAB, which was used for the

implementation. Afterwards, the EDA and ECG features were fused together into one matrix. The resulting matrix was 132,751 samples (rows) × 118 features (columns) (1 EDA reading + 1 ECG reading + 62 EDA features + 54 ECG features) in size. Other authors (see Section 2.3.3) operated on EDA and ECG separately, while feature fusion was applied to operate on both concurrently in this thesis.

4.1.3. Feature Preprocessing

During this work, it was noticed that some of RECOLA's physiological feature vectors contained missing values. A missing value can also be called not a number (NaN), which represents an undefined or unrepresentable value; for example, the result of dividing a number by zero. Missing values can negatively influence the performance of machine learning techniques. Hence, they were replaced with zeros.

Replacing missing values with zeros may introduce bias. However, very few missing values existed in the data. For example, only 2,080 samples out of 132,751 for one of the 118 physiological feature vectors, and 2,158 samples out of 132,751 for another (roughly 1.6%). The predesigned visual and acoustic feature sets did not include any missing values. Thus, the replacement of missing values would not have influenced the data to the point of yielding bias. Also, an outliers-based feature selection mechanism (see Section 4.1.4) was introduced to select relevant features and eliminate bias concerns if any were caused by the applied processing steps. The removal of missing values was necessary before training machine learning models. Otherwise, missing values would lead to training errors.

To study the impact on the regression performance, multiple data preprocessing techniques were investigated. In particular, standardizing/normalizing the EDA and ECG features using the z-score method was explored:

$$x_i' = z_i = \frac{x_i - \mu}{\sigma} \quad (5)$$

where x_i is a sample feature value; while μ and σ are the mean and the STD of the feature vector, respectively. This approach standardized the data in each feature vector to have a mean of 0 and an STD of 1. This step was performed using MATLAB's $zscore(X)$ method on each feature vector [13].

Furthermore, feature scaling was explored to normalize the range of all features between 0 and 1, for them to equally contribute to the prediction of the regressors. To bring all values into the range [0 1], feature scaling was applied on the EDA and ECG features through the min-max normalization approach [6464]:

$$x_i' = \frac{x_i - X_{min}}{X_{max} - X_{min}} \quad (6)$$

where X_{min} and X_{max} are the minimum and maximum values of a feature vector. The range of all features should be normalized for them to contribute proportionately to the final estimate of the regressors.

4.1.4. Feature Selection

In this study, the main sources of outliers in the features could be due to sensor malfunctions, human errors, noise from participant-specific behaviors, or natural variations between participants. A feature selection mechanism which depends on the number of outliers in a given feature vector was implemented. To be considered an outlier, the value of a feature should be "more than three scaled median absolute deviations from the median" [13]. The implemented feature selection method first detected outliers within each feature vector. Then, it checked if the number of outliers in any feature vector was higher than an acceptable percentage of the data. If so, the

corresponding feature vector was discarded. Only feature vectors that contained a number of outliers that was less than or equal to the acceptable percentage of the data were selected for machine learning. For instance, when the acceptable percentage of outliers was set to 5%, 50 features were selected as a result. When the acceptable percentage of outliers was set to 15%, 103 features were selected as a result. The more outliers were allowed, the more features passed the features selection test.

4.1.5. Data Shuffling and Splitting

Data shuffling is necessary to ensure the randomization and diversity of the data. The data were shuffled and split, where 80% were used for training and validation, and 20% were used for testing. The shuffled indexes were saved for use at later stages of this study. Since recordings were for different participants and time steps, the shuffled indexes were needed to ensure that the training, validation, and testing records were matching for all data modalities. Table 2 represents the breakdown of the data.

Table 2. Physiological Data Breakdown

Training Samples	Testing Samples	Total
106,201	26,550	132,751

4.1.6. Annotation Labelling of Physiological Samples

The data in RECOLA are labelled in the arousal and valence affective dimensions, and manually annotated using ANNEMO, a web slider-based annotation tool. Each recording was annotated by six native French speakers. The average of these six ratings was used to label the data in this study. These arousal and valence annotations are also sampled every 40 milliseconds. Again, as proceeded in Section 4.1.1, the first 50 annotations (2 seconds × 25 samples per second) were ignored. The remaining annotations were accordingly used to label the corresponding physiological, acoustic, and visual samples. All labelling and fusion of data samples and features were done based on the recordings time.

4.2. Regression over Physiological Data

Both classical and deep machine learning methods were exploited to perform continuous dimensional emotion predictions of arousal and valence values over physiological data. This work corresponds to stages 3 and 4 of this research (see Section 1.3).

4.2.1. Classical Regression

Four tree regressors (fine tree, medium tree, coarse tree, and optimizable tree), as well as ensembled regressors such as boosted trees, bagged trees, and an optimizable ensemble based on bagging and Bayesian optimization were trained and validated. 5-fold cross validation was performed during training to protect against overfitting. This partitions the data into 5 folds, where 4 folds were used for training and the remaining 1 fold was used for validation. For each fold, the model was trained on the data in the training folds, and the model's performance was assessed using the validation fold. This was repeated by alternating between the training and validation folds. Afterwards, the average validation error over all folds was calculated.

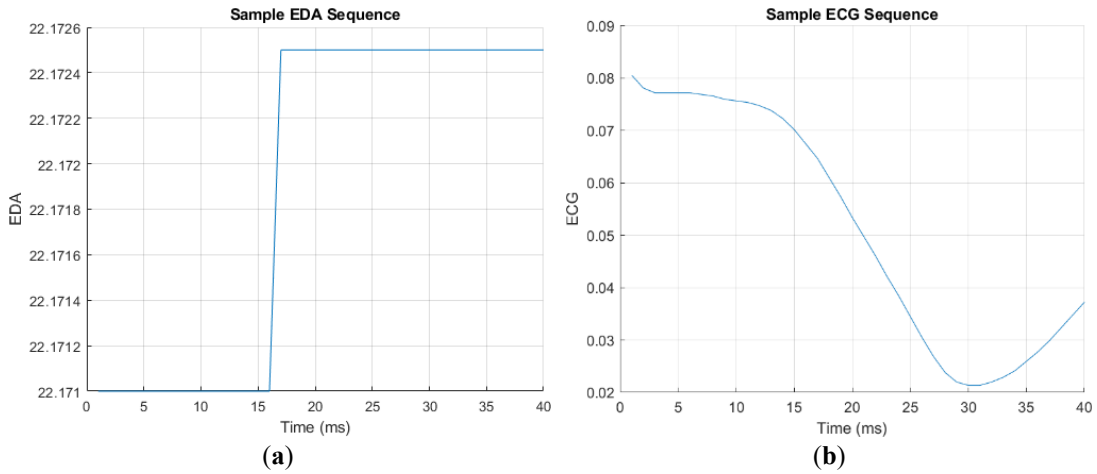
For the classical machine learning approach, RECOLA's EDA and ECG recordings along with their features were utilized, as described earlier. Table 3 displays the dimensions of the data sets used for classical regression. As discussed in 4.1.2, the baseline EDA and ECG features of RECOLA, described in [18], were fused. In [18], there were 62 EDA features, and 54 ECG features provided.

Table 3. Summary of Classical Machine Learning Data

Data set	Samples	EDA	ECG	Final Dimensions
Training	106,201	1 EDA reading+62 features	1 ECG reading+54 features	106,201×118
Validation		5-fold cross validation on training data		
Testing	26,550	1 EDA reading+62 features	1 ECG reading+54 features	26,550×118

4.2.2. Recurrent Neural Network (RNN) Regression

A BiLSTM RNN was trained and validated, using sequences of ECG and EDA, to predict arousal and valence values. As mentioned before, a BiLSTM network learns bidirectional long-term dependencies between time steps of time series or sequence data [13]. These dependencies can be useful when one wants the network to learn from the complete time series at each time step. In this case, the EDA and ECG signals were processed into sequences. Since the physiological signals contained in RECOLA were sampled every 1 millisecond, unlike the other modalities, which were sampled every 40 milliseconds, the physiological signal samples were grouped into 40-millisecond sequences. For instance, the sequence of physiological recordings from 0 to 39 milliseconds (inclusive) was labelled with the annotation at the 0^{th} millisecond, the sequence of physiological recordings from 40 to 79 milliseconds (inclusive) was labelled with the annotation at the 40^{th} millisecond, and so on. Figure 8 displays sample plots of these 40-millisecond sequences. As a result, 132,751 labelled sequences of EDA and ECG readings (without features) were acquired, each composed of 40 samples. That provided a total of 10,620,080 (132,751×40×2) physiological samples. The data sequences were divided into a training set of 106,201 sequences, and a testing set of 26,550 sequences. The training set was further broken by an 80-20% split to allow for validation. As a result, the final training set contained 84,960 sequences, and the validation set contained 21,241 sequences. It is important to mention that the feature sets were not needed for training the BiLSTM RNN model, since it is a deep learning approach, where the neural network learns features directly from the data. Thus, the processing steps from Sections 4.1.2, 4.1.3, and 4.1.4 were not applied on the data used for RNN regression.



(a) (b)

Figure 8. Sample a) EDA and b) ECG sequences.

Table 4 shows the training parameters of the BiLSTM network. Experimentally, the initial learning rate was set to 0.0001, and the number of epochs was set to 30. As there were 84,960 training sequences, the minimum batch size was set to 9 in order to evenly divide the training data into 9,440 equal batches and ensure the whole training set was used during each epoch. This resulted in 9,440 iterations per epoch (84,960/9 = 9,440). For validation frequency, the number of iterations was divided by 2 to ensure that the training process was validated at least twice per training epoch. The number of hidden units was set to 125. Since the network was bidirectional, the number of hidden units multiplied by 2; that is 250 units. The number of hidden units reflects the amount of information remembered between time steps [13]. The hidden state can contain information from all previous samples/time steps, irrespective of sequence length. A BiLSTM can include past and future information. Lastly, the stochastic gradient descent with momentum (SGDM) optimizer was used for training.

Table 4. BiLSTM Training Parameters

Parameter/Option	Amount/Value
Original Sequences	132,751
Training Sequences	84,960
Validation Sequences	21,241
Testing Sequences	26,550

Parameter/Option	Amount/Value
Sequence Length	40 milliseconds/samples
Learning Rate	0.0001
Minimum Batch Size	9
Number of Epochs	30
Iterations per Epoch	84,960/9 = 9,440
Validation Frequency	9,440/2 = 4,720
Hidden Units	125×2 = 250
Optimizer/Learner	SGDM

4.3. Results on Physiological Data

In this section, physiological features were processed through min-max scaling, z-score standardization, and outliers-based feature selection as explained in the earlier sections of this chapter. Then, regression models were trained on the processed physiological features. It was noticed that the aforementioned preprocessing steps had an influence on prediction performances. This work corresponds to stages 3 to 5, and 13 of this research and is related to contribution 5 (see Section 1.3). This section further presents the results that were obtained on physiological data from RECOLA.

4.3.1. Classical Regression Results

As described in Section 4.2.1, seven regression models including a fine regression tree, a medium regression tree, a coarse regression tree, and an optimizable tree were tested. Ensemble models such as boosted trees, bagged trees, and an optimizable ensemble were also tested. Table 5 summarizes the validation and testing prediction performances of these classical machine learning regressors in terms of the RMSE, PCC, and CCC performance measures, before feature processing.

In the tables, the validation results were computed through 5-fold cross-validation over the training data. The testing results were obtained by having the trained model predict arousal and valence values on the testing set. The row corresponding to the highest prediction performance, for each case, is displayed in bold font in the tables. For

detailed tables of all obtained prediction performances from the different regressors, please refer to 0.

Table 5. Classical Regression Prediction Performances before Replacing Missing Values and without Feature Standardization and Scaling (i.e., No Feature Processing)

Prediction	Scaled?	Standardized (z-score)?	Has NaNs?	Regressor	Validation RMSE	Testing RMSE, PCC, CCC
Arousal	No	No	Yes	Fine Tree	0.048101	0.0429, 0.9740, 0.9739
				Medium Tree	0.056231	0.0502, 0.9642, 0.9640
				Coarse Tree	0.077392	0.0698, 0.9292, 0.9276
				Optimizable Tree	0.046698	0.0405, 0.9769, 0.9769
				Boosted Trees	0.14376	0.1440, 0.6811, 0.5234
				Bagged Trees	0.028162	0.0233, 0.9941, 0.9918
				Optimizable Ensemble	**0.021198**	**0.0173, 0.9966, 0.9956**
Valence	No	No	Yes	Fine Tree	0.028757	0.0248, 0.9814, 0.9813
				Medium Tree	0.033463	0.0288, 0.9748, 0.9747
				Coarse Tree	0.047371	0.0429, 0.9431, 0.9422
				Optimizable Tree	0.027821	0.0243, 0.9821, 0.9821
				Boosted Trees	0.10128	0.1003, 0.6669, 0.4924
				Bagged Trees	0.018977	0.0156, 0.9943, 0.9921
				Optimizable Ensemble	**0.014697**	**0.0126, 0.9957, 0.9950**

According to Table 5, the optimizable ensemble regressor outperformed the other models. For arousal predictions, it achieved a testing RMSE, PCC, and CCC of 0.0173, 0.9966, and 0.9956, respectively. For valence predictions, it obtained a testing RMSE, PCC, and CCC of 0.0126, 0.9957, and 0.9950, respectively. These prediction performances are better than any prediction performances, reported on RECOLA, in other state-of-the-art studies as far as is known.

Figure 9 displays a plot of the predicted arousal and valence values against the actual values, as per the best model before feature processing. A perfect regression model predicts values equal to the actual values; in which case, all points would be on the diagonal line [13]. The vertical distance from the diagonal line to any point represents the prediction error for that point. Good models have small errors, where the predictions are scattered near the line.

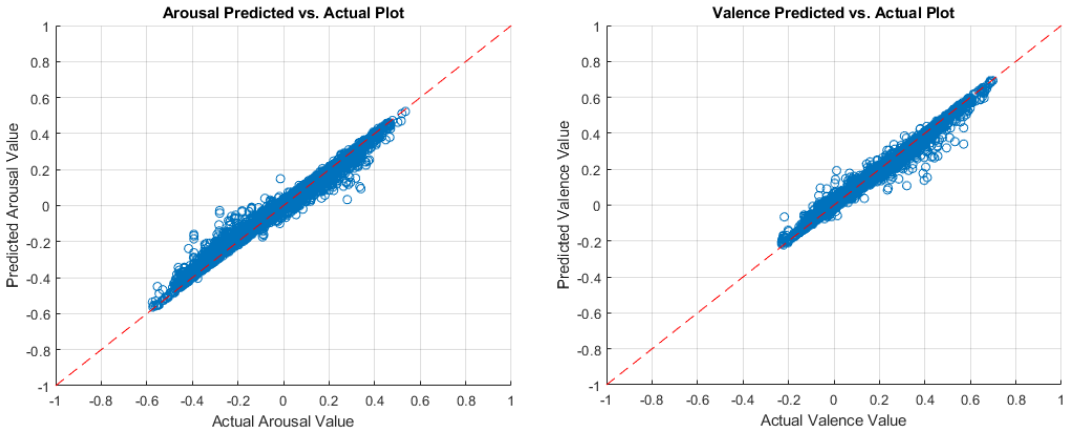


Figure 9. Plots of the predicted versus actual values of (a) arousal and (b) valence, without feature processing, as per the optimizable ensemble model.

Table 6 presents the validation and testing prediction performances of the classical machine learning regressors, after replacing missing values with zeros. The table shows that the replacement of missing values had influenced the prediction performances. For arousal, the best performance was attained through the bagged trees regressor. It reached a testing RMSE, PCC, and CCC of 0.0225, 0.9946, and 0.9925, respectively. This represents lower performance than the one reported in Table 5. For valence, the best performance was still achieved by the optimizable ensemble regressor. It obtained a testing RMSE, PCC, and CCC of 0.0105, 0.9970, and 0.9965, respectively. This shows an improvement in prediction performance over using the raw data with missing values. Therefore, it is better to replace missing values with zeros.

Table 6. Classical Regression Prediction Performances without Missing Values

Prediction	Scaled?	Standardized (z-score)?	Has NaNs?	Regressor	Validation RMSE	Testing RMSE, PCC, CCC
Arousal	No	No	No	Fine Tree	0.047736	0.0420, 0.9751, 0.9751
				Medium Tree	0.055758	0.0494, 0.9653, 0.9651
				Coarse Tree	0.076991	0.0697, 0.9296, 0.9279
				Optimizable Tree	0.045999	0.0396, 0.9780, 0.9779
				Boosted Trees	0.14362	0.1440, 0.6821, 0.5237
				Bagged Trees	**0.027193**	**0.0225, 0.9946, 0.9925**
				Optimizable Ensemble	0.055619	0.0532, 0.9596, 0.9584
Valence	No	No	No	Fine Tree	0.026639	0.0237, 0.9829, 0.9829
				Medium Tree	0.032055	0.0281, 0.9760, 0.9760
				Coarse Tree	0.046648	0.0422, 0.9449, 0.9442
				Optimizable Tree	0.025588	0.0230, 0.9840, 0.9840
				Boosted Trees	0.099214	0.0987, 0.6784, 0.5164
				Bagged Trees	0.016937	0.0136, 0.9959, 0.9940
				Optimizable Ensemble	**0.012208**	**0.0105, 0.9970, 0.9965**

Figure 10 shows a plot of the predicted arousal and valence values against the actual values, as per the best model after the removal of missing values. In Figure 10, the valence predictions are more spread along the diagonal line, showing a better performance over Figure 9. On the other hand, more of the arousal predictions appear farther away from the diagonal line than in Figure 9.

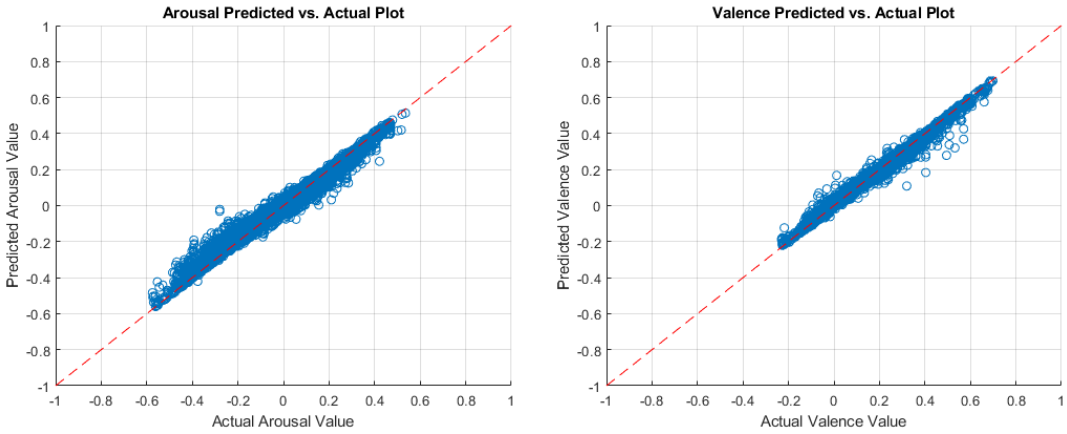


Figure 10. Plots of the predicted versus actual values of (a) arousal and (b) valence, after the removal of missing values, as per the bagged trees and optimizable ensemble models, respectively.

Feature scaling did not enhance the prediction performance. Table 7 summarizes the validation and testing prediction performances of the classical machine learning regressors, after replacing missing values with zeros, applying z-score feature standardization/normalization, and scaling the feature set via min-max normalization. The best performance was attained by the bagged trees regressor for arousal prediction, with a testing RMSE, PCC, and CCC of 0.0225, 0.9946, and 0.9925, respectively. This represents lower performance than Table 5, but the same as Table 6. The best performance was still reached via the optimizable ensemble regressor for valence predictions, with a testing RMSE, PCC, and CCC of 0.0115, 0.9965, and 0.9958, respectively. This shows lower performance than Table 6, but better than Table 5.

Table 7. Classical Regression Prediction Performances after Feature Scaling Only (No Missing Values)

Prediction	Scaled?	Standardized (z-score)?	Has NaNs?	Regressor	Validation RMSE	Testing RMSE, PCC, CCC
Arousal	Yes	No	No	Fine Tree	0.047711	0.0420, 0.9751, 0.9750
				Medium Tree	0.055753	0.0495, 0.9652, 0.9651
				Coarse Tree	0.076989	0.0697, 0.9296, 0.9279

Prediction	Scaled?	Standardized (z-score)?	Has NaNs?	Regressor	Validation RMSE	Testing RMSE, PCC, CCC
				Optimizable Tree	0.045976	0.0396, 0.9779, 0.9779
				Boosted Trees	0.14362	0.1440, 0.6821, 0.5237
				Bagged Trees	**0.027009**	**0.0225, 0.9946, 0.9925**
				Optimizable Ensemble	0.027166	0.0237, 0.9926, 0.9918
Valence	Yes	No	No	Fine Tree	0.027735	0.0237, 0.9829, 0.9829
				Medium Tree	0.032494	0.0281, 0.9760, 0.9760
				Coarse Tree	0.046696	0.0422, 0.9449, 0.9442
				Optimizable Tree	0.026567	0.0230, 0.9840, 0.9840
				Boosted Trees	0.099295	0.0987, 0.6784, 0.5164
				Bagged Trees	0.016949	0.0137, 0.9958, 0.9940
				Optimizable Ensemble	**0.014091**	**0.0115, 0.9965, 0.9958**

Figure 11 displays a plot of the predicted arousal and valence values against the actual values, as per the best model after the removal of missing values and feature scaling.

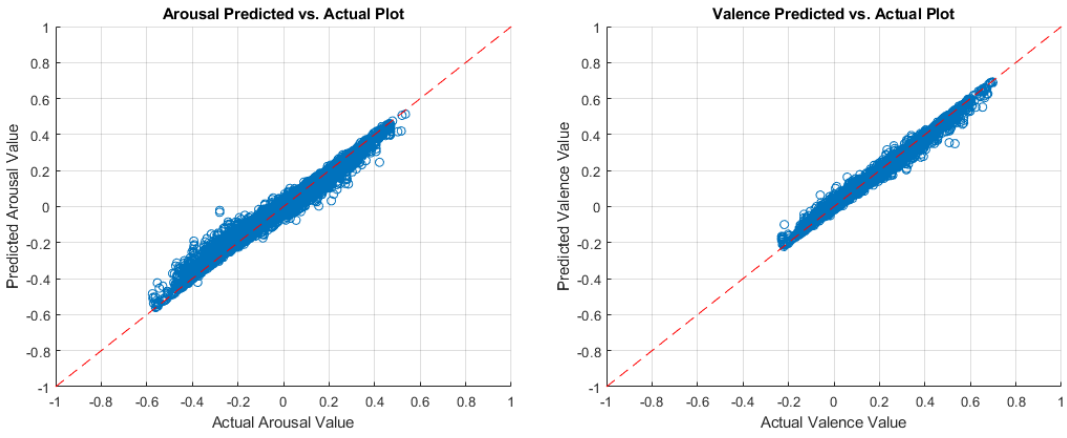


Figure 11. Plots of the predicted versus actual values of (a) arousal and (b) valence, after the removal of missing values and feature scaling, as per the bagged trees and optimizable ensemble models, respectively.

Feature standardization/normalization through the z-score method had positively influenced the prediction performance of both arousal and valence predictions. Table 8 summarizes the validation and testing prediction performances of the classical machine learning regressors, after replacing missing values with zeros and applying z-score feature standardization/normalization. For both arousal and valence, the best performance was achieved via the optimizable ensemble regressor. It obtained a testing RMSE, PCC, and CCC of 0.0170, 0.9967, and 0.9958 at arousal predictions and 0.0106, 0.9971, and 0.9965 at valence predictions, respectively. These prediction performances outperform the prediction performances from Tables 5-7.

Table 8. Classical Regression Prediction Performances after Feature Standardization Alone (No Missing Values)

Prediction	Scaled?	Standardized (z-score)?	Has NaNs?	Regressor	Validation RMSE	RMSE, PCC, CCC
Arousal	No	Yes	No	Fine Tree	0.046975	0.0420, 0.9751, 0.9751
				Medium Tree	0.055288	0.0494, 0.9653, 0.9651
				Coarse Tree	0.076615	0.0697, 0.9296, 0.9279
				Optimizable Tree	0.045176	0.0396, 0.9780, 0.9779
				Boosted Trees	0.14359	0.1440, 0.6821, 0.5237
				Bagged Trees	0.02689	0.0221, 0.9948, 0.9927
				Optimizable Ensemble	**0.021043**	**0.0170, 0.9967, 0.9958**
Valence	No	Yes	No	Fine Tree	0.026728	0.0237, 0.9830, 0.9830
				Medium Tree	0.031906	0.0280, 0.9761, 0.9760
				Coarse Tree	0.046718	0.0422, 0.9450, 0.9443
				Optimizable Tree	0.025805	0.0230, 0.9840, 0.9840
				Boosted Trees	0.099252	0.0987, 0.6784, 0.5164
				Bagged Trees	0.016891	0.0137, 0.9958, 0.9940

Prediction	Scaled?	Standardized (z-score)?	Has NaNs?	Regressor	Validation RMSE	RMSE, PCC, CCC
				Optimizable Ensemble	**0.012907**	**0.0106, 0.9971, 0.9965**

Figure 12 displays a plot of the predicted arousal and valence values against the actual values, as per the best model after the removal of missing values and feature standardization.

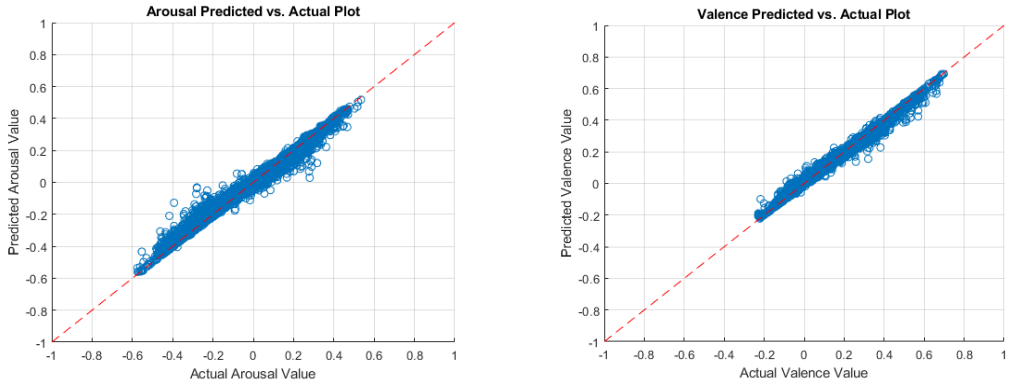


Figure 12. Plots of the predicted versus actual values of (a) arousal and (b) valence, after the removal of missing values and feature standardization, as per the optimizable ensemble model.

Table 9 summarizes the validation and testing prediction performances of the classical machine learning regressors, after replacing missing values with zeros as well as applying both feature scaling and z-score feature standardization/normalization. The table shows that combining feature scaling and standardization had positively influenced the prediction performance of valence predictions. That was not the case for arousal predictions, however. The best performance was attained via the optimizable ensemble regressor. For arousal, it reached a testing RMSE, PCC, and CCC of 0.0194, 0.9954, and 0.9945, respectively. For valence, it achieved a testing RMSE, PCC, and CCC of 0.0104, 0.9971, and 0.9966, respectively. As such, one can conclude that a processing mechanism which includes replacing missing values and z-score standardization produces the best prediction performance at predicting the emotional measure of arousal. On the other hand, a processing mechanism which includes

replacing missing values, feature scaling, and z-score standardization produces the best prediction performance at predicting the emotional measure of valence.

Table 9. Classical Regression Prediction Performances after BOTH Feature Scaling and Standardization (No Missing Values)

Prediction	Scaled?	Standardized (z-score)?	Has NaNs?	Regressor	Validation RMSE	RMSE, PCC, CCC
Arousal	Yes	Yes	No	Fine Tree	0.04771	0.0420, 0.9751, 0.9750
				Medium Tree	0.05575	0.0494, 0.9653, 0.9651
				Coarse Tree	0.076989	0.0697, 0.9296, 0.9279
				Optimizable Tree	0.045975	0.0396, 0.9779, 0.9779
				Boosted Trees	0.14362	0.1440, 0.6821, 0.5237
				Bagged Trees	0.026945	0.0227, 0.9945, 0.9923
				Optimizable Ensemble	**0.023379**	**0.0194, 0.9954, 0.9945**
Valence	Yes	Yes	No	Fine Tree	0.027736	0.0237, 0.9830, 0.9830
				Medium Tree	0.032494	0.0281, 0.9760, 0.9760
				Coarse Tree	0.046696	0.0422, 0.9449, 0.9442
				Optimizable Tree	0.026569	0.0229, 0.9841, 0.9841
				Boosted Trees	0.099296	0.0987, 0.6784, 0.5164
				Bagged Trees	0.016800	0.0135, 0.9959, 0.9942
				Optimizable Ensemble	**0.012570**	**0.0104, 0.9971, 0.9966**

Figure 13 displays a plot of the predicted arousal and valence values against the actual values, as per the best model after the removal of missing values and feature scaling and standardization.

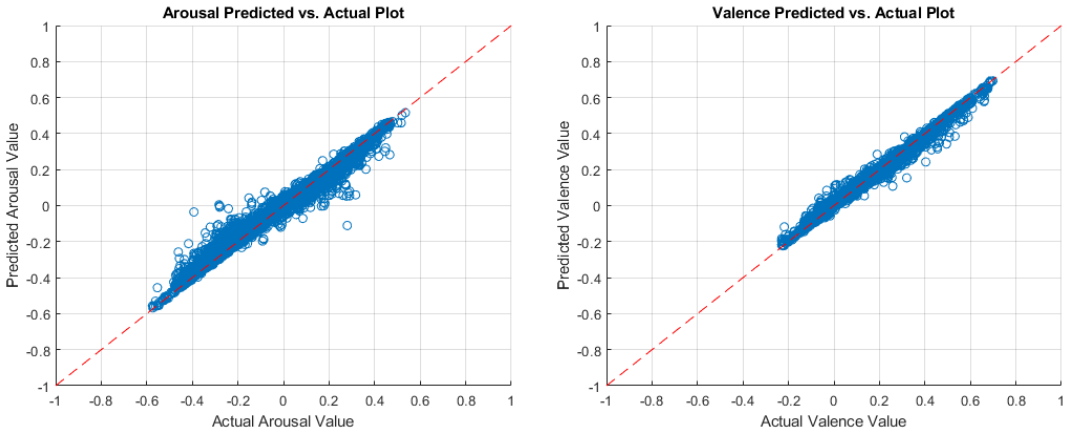


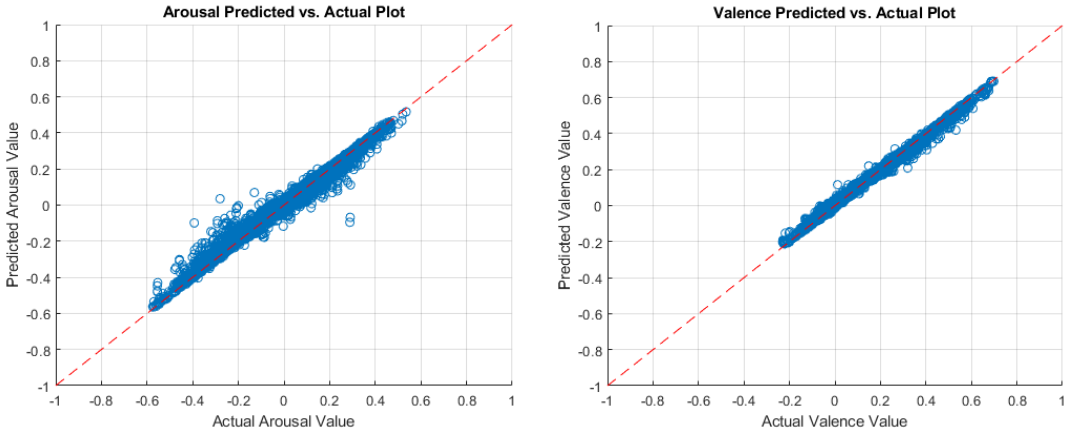
Figure 13. Plots of the predicted versus actual values of (a) arousal and (b) valence, after the removal of missing values, feature scaling and standardization, as per the optimizable ensemble model.

In an attempt to further improve the prediction performance, a novel outliers-based feature selection mechanism was proposed. After feature selection, the prediction performance improved. A feature selection mechanism which allows only 5% of outliers (50 features) produced better prediction performances for arousal predictions. On the other hand, allowing up to 15% of outliers (103 features) produced better prediction performances for valence predictions. With 50 features selected, the optimizable ensemble regressor obtained an RMSE, PCC, and CCC of 0.0168, 0.9965, and 0.9959 when predicting arousal values. With 103 features selected, the optimizable ensemble regressor attained an RMSE, PCC, and CCC of 0.0083, 0.9985, and 0.9978 when predicting valence values.

Table 10 summarizes the best validation and testing prediction performances of the classical machine learning regressors, after feature scaling, standardization, and selection. Figure 14 displays a plot of the predicted arousal and valence values against the actual values, as per the best model after all feature processing steps.

Table 10. Classical Regression Prediction Performances after Feature Scaling, Standardization, and Selection

Prediction	Regressor	Acceptable Outliers	Selected Features	Validation RMSE	RMSE, PCC, CCC
Arousal	**Optimizable Ensemble**	**5%**	**50**	**0.020137**	**0.0168, 0.9965, 0.9959**
	Optimizable Ensemble	15%	103	0.054835	0.0533, 0.9597, 0.9574
Valence	Optimizable Ensemble	5%	50	0.012644	0.0099, 0.9976, 0.9969
	Optimizable Ensemble	**15%**	**103**	**0.010682**	**0.0083, 0.9985, 0.9978**



(a) **(b)**

Figure 14. Plots of the predicted versus actual values of (a) arousal and (b) valence, after all feature processing steps, as per the optimizable ensemble model.

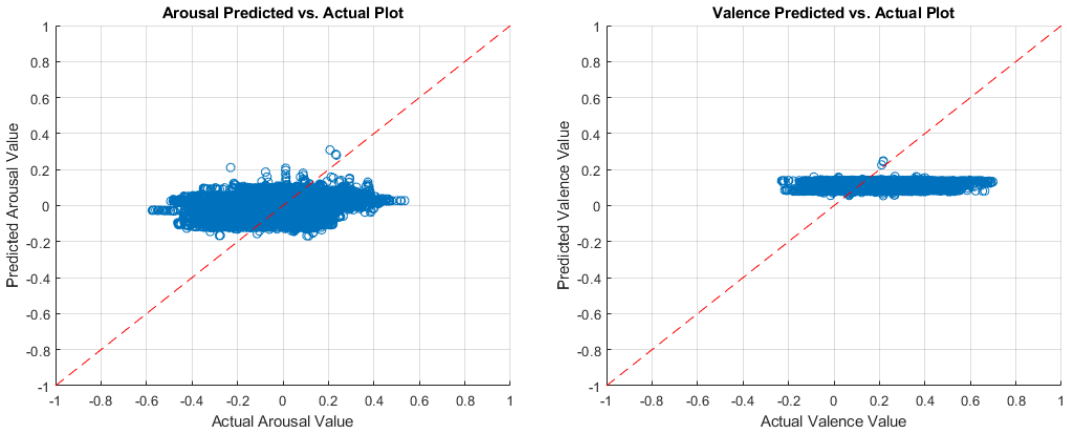
4.3.2. Recurrent Neural Network (RNN) Results

The trained BiLSTM RNN was tested with a data set of 26,550 sequences. Each sequence was 40 milliseconds/samples in length. Table 11 summarizes the validation and testing prediction performances from BiLSTM regression in terms of the RMSE, PCC, and CCC performance measures. The validation results were acquired through 80-20% split validation. The testing results were acquired by having the trained model predict the arousal and valence values of the testing sequences.

Table 11. BiLSTM Regression Prediction Performances

Prediction	Validation RMSE	Testing RMSE, PCC, CCC
Arousal	0.18461	0.1832, 0.2515, 0.1089
Valence	0.12774	0.1268, 0.1773, 0.0567

Figure 15 displays a plot of the predicted arousal and valence values against the actual values, as per the BiLSTM model.



(a) **(b)**

Figure 15. Plots of the predicted versus actual values of (a) arousal and (b) valence as per the BiLSTM model.

Table 11 shows that the classical regression prediction performances are better than the prediction performances of the BiLSTM model. However, the attained prediction performances are comparable to the prediction performances obtained in [29] and outperform the baseline prediction performances from [26]. For arousal, the achieved prediction performances are better than those of [29], but that is not the case for valence on the other hand.

In conclusion, Table 12 compares the best reached prediction performances with the baseline prediction performances [26] as well as the prediction performances of [29]. The table shows the prediction performances that have been reached through the optimizable ensemble regressor, after the application of min-max scaling, z-score

standardization, and selection on physiological features. Only 50 features were selected for arousal predictions, whereas 103 features were selected for valence predictions.

Table 12. Comparison of RNN Prediction Performances

Prediction	Input	RMSE	PCC	CCC
Arousal	**Optimizable Ensemble**	**0.0168**	**0.9965**	**0.9959**
	EDA + ECG BiLSTM	0.183	0.252	0.109
	EDA LSTM [29]	0.250	0.089	0.082
	ECG LSTM [29]	0.218	0.407	0.357
	EDA Baseline [26]	-	-	0.029
	ECG Baseline [26]	-	-	0.065
Valence	**Optimizable Ensemble**	**0.0083**	**0.9985**	**0.9978**
	EDA + ECG BiLSTM	0.127	0.177	0.057
	EDA LSTM [29]	0.124	0.267	0.177
	ECG LSTM [29]	0.117	0.412	0.364
	EDA Baseline [26]	-	-	0.058
	ECG Baseline [26]	-	-	0.043

4.3.3. Unimodal Decision Fusion Results

In an attempt to enhance the prediction performances, the predictions from multiple regressors were fused to see if further improvement can be achieved. As such, initially, the predictions of the tree regressors were fused together, and of the ensembled regressors together. Then, the predictions of all models were also combined together. Table 13 shows the prediction performances after feature scaling, feature standardization, and decision fusion (i.e., without feature selection).

Table 13. Decision Fusion Prediction Performances before Feature Selection

Prediction	Scaled?	Standardized (z-score)?	Has NaNs?	Fused Models	RMSE, PCC, CCC
Arousal	Yes	Yes	No	**Optimizable Ensemble**	**0.0194, 0.995, 0.995**
				Fine, Medium, Coarse, and Optimizable Trees	0.0409, 0.976, 0.976
				Boosted Trees, Bagged Trees, Optimizable Ensemble	0.0571, 0.981, 0.941
				All Trees, All Ensembles, BiLSTM	0.0534, 0.984, 0.950

Prediction	Scaled?	Standardized (z-score)?	Has NaNs?	Fused Models	RMSE, PCC, CCC
Valence	Yes	Yes	No	**Optimizable Ensemble**	**0.0104, 0.997, 0.997**
				Fine, Medium, Coarse, and Optimizable Trees	0.0241, 0.982, 0.982
				Boosted Trees, Bagged Trees, Optimizable Ensemble	0.0381, 0.983, 0.944
				All Trees, All Ensembles, BiLSTM	0.0346, 0.988, 0.955

Table 14 shows the prediction performances after feature scaling, feature standardization, feature selection, and decision fusion. As can be seen, unimodal decision fusion did not outperform the performance that was previously observed using the optimizable ensemble method alone.

Table 14. Decision Fusion Prediction Performances after Feature Selection

Prediction	Feature Selection	Fused Models	RMSE, PCC, CCC
Arousal	5% Outliers	**Optimizable Ensemble**	**0.0168, 0.997, 0.996**
		Fine, Medium, Coarse, and Optimizable Trees	0.0386, 0.979, 0.978
		Boosted Trees, Bagged Trees, Optimizable Ensemble	0.0567, 0.980, 0.943
		All Trees, All Ensembles, BiLSTM	0.0426, 0.983, 0.971
Valence	15% Outliers	**Optimizable Ensemble**	**0.0083, 0.999, 0.998**
		Fine, Medium, Coarse, and Optimizable Trees	0.0236, 0.983, 0.983
		Boosted Trees, Bagged Trees, Optimizable Ensemble	0.0375, 0.983, 0.946
		All Trees, All Ensembles, BiLSTM	0.0276, 0.985, 0.973

In this section, decision fusion was applied on the predictions that were obtained on all physiological features as well as on selected physiological features. It was noticed that decision fusion has not improved the prediction performances over the optimizable ensemble alone. However, the prediction performances that were obtained via decision fusion are much better than the literature. This work corresponds to stage 5 and 13 of this research and is related to contribution 5 (see Section 1.3). The best prediction performances on physiological data in the literature were achieved by the authors of [29]. For arousal, they obtained an RMSE of 0.218, a PCC of 0.407, and a CCC of

0.357. For valence, they attained an RMSE of 0.117, a PCC of 0.412, and a CCC of 0.364. In this study, the reached RMSE was 0.0168, PCC was 0.997, and CCC was 0.996 on arousal predictions. On valence predictions, the achieved RMSE was 0.0083, PCC was 0.999, and CCC was 0.998.

5. PROOF-OF-CONCEPT FOR AFFECT RECOGNITION USING VISUAL DATA

This chapter presents stages 2, 6 to 10, and 13 of this research and is related to contributions 1, 2, 3 to 5 (see Section 1.3). As in the case of physiological data, 18 RECOLA videos were preprocessed. All videos were preprocessed by applying frame extraction and sequencing, face detection and cropping, feature extraction, data shuffling and splitting, annotation labelling, and data augmentation, as detailed in the following section. After processing, the extracted images (i.e., video frames) of participants' faces, or half faces for VR applications – where the eyes are often hidden by an HMD (see Chapter 6) were inputted into CNNs for predicting arousal and valence values. The trained CNNs were then used to extract deep visual features. The predesigned visual features of RECOLA were also extracted and used as a proof-of-concept for predicting arousal and valence values via classical regression techniques. Figure 16 illustrates the proposed approach for visual data.

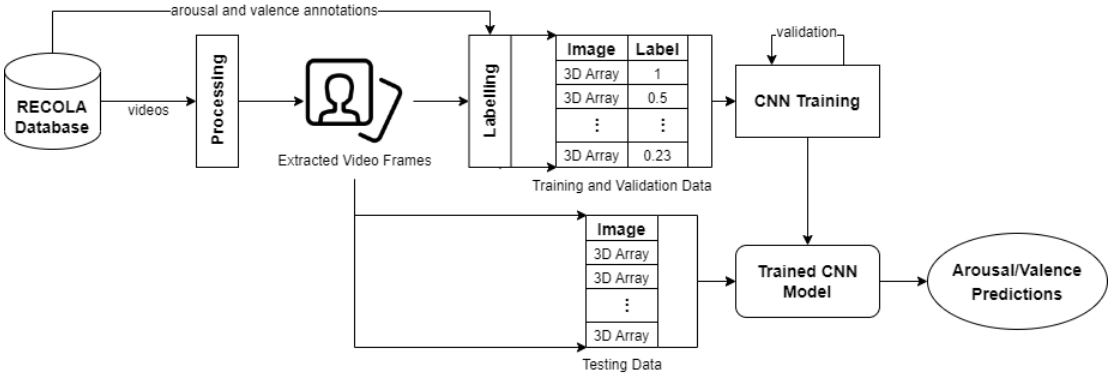


Figure 16. An overview of the proposed approach for visual data.

5.1. Processing of Visual Data

This section details the processing steps that were applied to the video recordings of RECOLA. Figure 17 outlines these steps, which are contained in the "Processing" block of Figure 16. This work corresponds to stages 2 and 7 of this research and is related to contribution 1 (see Section 1.3).

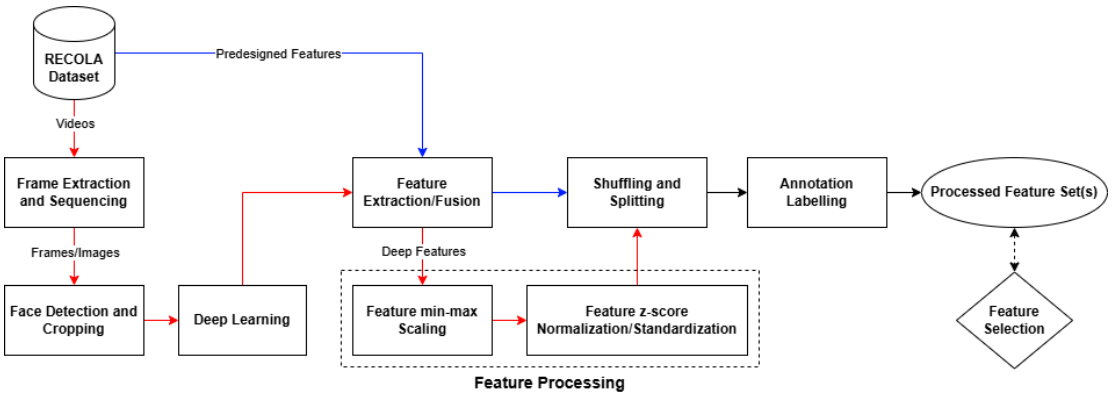


Figure 17. Breakdown of processing steps for visual data. The blue lines outline the flow for predesigned visual features, whereas the red lines outline the flow for the deep visual features. The black lines represent mutual steps.

5.1.1. Frame Extraction and Sequencing

The videos available in the RECOLA data set are approximately 5 minutes long each. They were processed by extracting their video frames at a rate of 25 frames per second. As a result, an image frame was obtained every 40 milliseconds of video recording. That is a total of approximately 7500 frames per video. To match the physiological data, the first 50 frames were skipped again, ensuring that the data were synchronous (i.e., had all samples for the same times of recording). For example, EDA and ECG readings, EDA and ECG features, and a video frame would be collected at the 40^{th} millisecond of recording, the 80^{th} millisecond of recording, and so on. The same shuffling indexes used for shuffling the physiological data were used here to randomize and shuffle the video frames to ensure data coherence. To facilitate the training,

validation, and testing process of CNNs, the visual data were then split using an 80-20% split, where 80% of the frames were used for training and validation, while 20% of them were used for testing. Table 15 represents the breakdown of the visual data.

Table 15. Visual Data Breakdown

Parameter	Original	80-20% Split
Training Frames	106,201	84,960
Validation Frames	N/A	21,241
Testing Frames	26,550	26,550
TOTAL	132,751	132,751

An image datastore object was then used to manage the collected images, and their corresponding arousal and valence labels. The datastore created a data structure that accessed image data from a specific location (e.g., on the computer). The references to the image files were later used to train CNNs (see Section 5.2.2).

5.1.2. Predesigned Visual Features

The video recordings of RECOLA were sampled at a sampling rate of 25 frames/s, where visual features were extracted for each video frame [23]. As predesigned visual features, RECOLA contains 20 attributes alongside their first order derivative, resulting in 40 features in total. These attributes/features include 15 facial AUs of emotional expressions, the head-pose in three dimensions (i.e., X, Y, Z), and the mean and STD of the optical flow in the region around the head. The AUs are AU1 (Inner Brow Raiser), AU2 (Outer Brow Raiser), AU4 (Brow Lowerer), AU5 (Upper Lid Raiser), AU6 (Cheek Raiser), AU7 (Lid Tightener), AU9 (Nose Wrinkler), AU11 (Nasolabial Deepener), AU12 (Lip Corner Puller), AU15 (Lip Corner Depressor), AU17 (Chin Raiser), AU20 (Lip Stretcher), AU23 (Lip Tightener), AU24 (Lip Pressor), and AU25 (Lips Part) from FACS. For more information about these features and their extraction, please refer to [23]. These features were used in this work to predict arousal and valence values via classical regression.

The visual features were calculated on each frame, meaning that they were provided every 40 milliseconds as well. Since other data modalities of RECOLA only started being recorded after 2 second (2000 milliseconds), any readings that occurred before that time were skipped. As a result, the first 50 frames (2 second × 25 frames/second) of the recordings were unused in this work.

The predesigned features were also shuffled and split, where 80% went towards training and validation, and 20% went towards testing. The training and validation data set for classical regression was 106201 frames × 40 features in size, while the testing data set was 26550 frames × 40 features in size. These data sets were later used to train and test classical regressors (see Section 5.2.1).

5.1.3. Face Detection and Cropping

Face detection was then applied to narrow the prediction area of the arousal and valence emotional measures. A cascade object detector was utilized to detect faces within the video frames. The cascade object detector is based on the Viola-Jones algorithm to detect people's faces; face parts like noses, eyes, mouths; or upper bodies [13,65]. The cascade object detector locates faces in images by a sliding window across the image. Then, it uses a cascade classifier to decide if the window includes a face. The cascade classifier is comprised of stages containing ensembles of weak learners. The learners are decision stump classifiers, where each stage is trained using the boosting technique. Boosting enables the training of highly accurate classifiers by taking a weighted mean of the decisions of the weak learners. The size of the sliding window adjusts to detect faces at various scales, whereas aspect ratio of the window stays the same. The cascade object detector is very sensitive to out-of-plane rotation, since rotations change the aspect ratio of most 3D objects. Hence, a detector for each orientation of the face is trained for better detection results.

Each stage of the classifier labels the region defined by the current location of the sliding window as either positive or negative. Positive indicates that an object was found and negative indicates no objects were found. If the label is negative, the classification of this region is complete, and the detector slides the window to the next location. If the label is positive, the classifier passes the region to the next stage. The detector reports an object found at the current window location when the final stage classifies the region as positive.

Following face detection, it was noticed that the algorithm failed to detect faces in some of the obtained video frames. Hence, these images were cropped according to the face coordinates of the nearest image with a detected face. In the best-case scenario, the nearest image with a detected face would be the image preceding or following the image with a missed face. In the worst-case scenario, the algorithm would have failed to detect faces in a group of images, where the nearest image with a detected face would be more than one video frame away. In this case, the coordinates of the face might be off due to the movement of the participant in the video. Thus, manual intervention to edit the images was required. Figure 18 displays a sample face image before and after face detection.

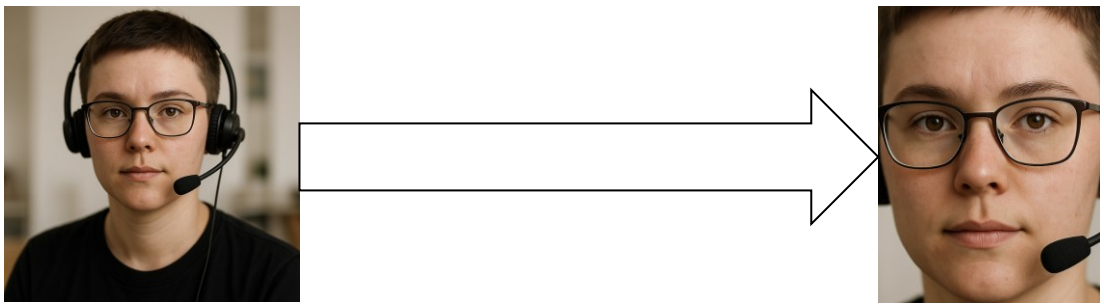


Figure 18. AI-generated face image, simulating a video frame before and after face detection.

5.1.4. Annotation Labelling of Visual Samples

Similar to the approach for physiological data, the first 50 annotations (2 seconds × 25 samples per second) were ignored. The remaining annotations were accordingly used to label the corresponding visual samples. All labelling and fusion of data samples and features were carried out based on the recordings time.

5.2. Regression over Visual Data

Both classical and deep machine learning methods were exploited to perform continuous dimensional emotion predictions of arousal and valence values over visual data. This work corresponds to stages 6, 8, and 9 of this research as presented in Section 1.3.

5.2.1. Classical Regression

For the prediction of arousal and valence values from the predesigned visual feature vectors, an optimizable ensemble regressor based on bagging and Bayesian optimization was used. Also, other regression models were experimented with for comparison purposes: tree regressors, regression kernels, and ensemble regression. In addition to the optimizable ensemble model, four tree regressors (fine, medium, coarse, and optimizable tree), two regression kernels (SVM and least squares regression kernel), and two ensembled regressors (boosted and bagged trees) were trained and validated. 5-fold cross-validation was implemented during training to avoid overfitting.

5.2.2. Convolutional Neural Network (CNN) Regression

Two pretrained CNNs were experimented with: ResNet-18 and MobileNet-v2. ResNet-18 is a CNN that is 18-layers-deep, whereas MobileNet-v2 has a depth of 53 layers [13]. Both of these pretrained CNNs are capable of classifying images of 1000 object

categories, such as a keyboard, mouse, pencil, and many animals. As a result, these networks have learned rich feature representations for a wide range of images.

Coloured images are read as 3D matrices/arrays of 8-bit or 16-bit unsigned integers in the format of rows × columns × RGB components. Each element in the rows and columns dimensions of the matrix corresponds to a single pixel in the displayed image. The first, second, and third planes in the third dimension of the matrix represent the red, green, and blue pixel intensities, respectively. ResNes-18 and MobileNet-v2 have an image input size of 224 × 224 × 3.

To fine-tune the pretrained CNNs for regression to predict arousal and valence values, the layers of each CNN were customized to suit the needs of this study and data augmentation was applied. Thus, the image input layer was replaced to make it accept images of size 280 × 280 × 3. Additionally, the final fully connected layer and the classification output layer were replaced with a fully connected layer of size 1 (the number of responses, i.e., arousal/valence value) and a regression layer. The convolutional layers of the CNNs extract image features that are then used by the last learnable layer and the final classification layer to classify the input image [13]. These layers have information about converting the extracted features into class probabilities, loss values, and predicted labels. In the cases of ResNet-18 and MobileNet-v2, the last learnable layer is the fully connected layer. The learning rates of the last learnable layer were adjusted in order to make the CNNs learn faster in the new fully connected layer than in the transferred/pretrained convolutional layers. This was done by setting the learning rate factors for weights and biases to 10.

The amount of training data was increased by applying randomized data augmentation. Data augmentation allows CNNs to train to be invariant to distortions in image data and helps to prevent overfitting, by preventing the CNN from memorizing the exact

characteristics of training images. Augmentation options such as random reflection in the *x*-axis, random rotation, and random rescaling were used. As aforementioned, the image input layer of the pretrained CNNs (ResNet-18 and MobileNet-v2) was replaced to allow them to take larger input images of size 280 × 280 × 3, but the images in the video frames if this study did not all have this size. Therefore, an augmented image datastore was used to automatically resize the images. Also, additional augmentation operations were specified to be performed on the images in order to prevent the CNNs from memorizing image features. The images were randomly reflected along the vertical *x*-axis, randomly rotated from the range [–90, 90] degrees, and randomly rescaled from the range [1, 2]. These changes did not affect the contents of the training images; however, they helped the CNNs in extracting/learning more features from the images.

The training options and parameters were modified depending on the size of the input data. Table 16 summarizes the training parameters that were used for training the CNNs.

Table 16. CNN Training Parameters

Parameter/Option	Amount/Value
Training Images	84,960
Validation Images	21,241
Testing Images	26,550
Learning Rate	0.0001
Minimum Batch Size	9
Number of Epochs	30
Iterations per Epoch	84,960/9 = 9440
Validation Frequency	9440/2 = 4720
Optimizer/Learner	SGDM

Experimentally, the initial learning rate was set to 0.0001, and the number of epochs was set to 30. As there were 84,960 training images, the minimum batch size was set to 9 in order to evenly divide the training data into 9,440 equal batches and ensure the whole training set was used during each epoch. This resulted in 9,440 iterations per epoch (84,960/9 = 9,440). For validation frequency, the number of iterations was

divided by 2 to ensure that the training process was validated at least twice per training epoch. The SGDM optimizer was used for training.

5.3. Results on Visual Data

In this section, the regression models were trained on the processed visual data. This work corresponds to stages 6, 8, 9, and 13 of this research and is related to contribution 5 (see Section 1.3). As described in Section 5.2.1, nine classical regressors: fine tree, medium tree, coarse tree, optimizable tree, SVM kernel, least squares regression kernel, boosted trees, bagged trees, and optimizable ensemble were tested for predicting arousal and valence values from visual features. Table 17 summarizes the validation and testing prediction performances of these regression models. The attained testing RMSE, PCC, and CCC were 0.1033, 0.8498, and 0.8001 on arousal predictions, respectively. The reached testing RMSE, PCC, and CCC were 0.07016, 0.8473, and 0.8053 on valence predictions, respectively. These performances were obtained using an optimizable ensemble regressor. They are better than those from the literature [23,26-36], who performed more complex processing and feature extraction.

Table 17. Optimizable Ensemble Prediction Performances on Predesigned Visual Features

Prediction	Regression Type	Validation RMSE	Testing RMSE, PCC, CCC
Arousal	Fine Tree	0.15389	0.1477, 0.6812, 0.6805
	Medium Tree	0.14601	0.1410, 0.6902, 0.6838
	Coarse Tree	0.14477	0.1410, 0.6731, 0.6516
	Optimizable Tree	0.14351	0.1396, 0.6861, 0.6719
	SVM Kernel	0.13665	0.1354, 0.7018, 0.6807
	Least Squares Kernel	0.13444	0.1331, 0.7097, 0.6633
	Boosted Trees	0.161	0.1607, 0.5463, 0.3743
	Bagged Trees	0.11285	0.1082, 0.8304, 0.7796
	Optimizable Ensemble	**0.10791**	**0.1033, 0.8498, 0.8001**
Valence	Fine Tree	0.10191	0.0981, 0.6975, 0.6967
	Medium Tree	0.097111	0.0944, 0.7011, 0.6947
	Coarse Tree	0.097623	0.0948, 0.6826, 0.6610
	Optimizable Tree	0.096525	0.0945, 0.6922, 0.6801

Prediction	Regression Type	Validation RMSE	Testing RMSE, PCC, CCC
	SVM Kernel	0.094882	0.0943, 0.6855, 0.6495
	Least Squares Kernel	0.092417	0.0916, 0.7030, 0.6574
	Boosted Trees	0.11142	0.1104, 0.5525, 0.3467
	Bagged Trees	0.074689	0.0714, 0.8421, 0.7962
	Optimizable Ensemble	**0.073335**	**0.0702, 0.8473, 0.8053**

As described in Section 5.2.2, two CNN models: ResNet-18 and MobileNet-v2 were tested for predicting arousal and valence values from images of faces. Table 18 summarizes the validation and testing prediction performances of these CNNs with images of the whole face. The obtained prediction performances outperform the baseline prediction performances, reported in [26], of 0.312 and 0.438 CCCs for arousal and valence, respectively. Additionally, they are comparable to other prediction performances from the literature (see Section 2.3.3).

Table 18. CNN Regression Prediction Performances on Images of Full Faces

CNN	Prediction	Number of Epochs	Validation RMSE	RMSE, PCC, CCC
ResNet-18	Arousal	30	0.15281	0.1524, 0.5952, 0.5446
		60	0.14386	0.1443, 0.6844, 0.6353
		150	0.12683	0.1284, 0.7605, 0.7336
	Valence	30	0.11651	0.1167, 0.6720, 0.5571
		150	0.08602	0.0855, 0.7669, 0.7403
MobileNet-v2	Arousal	30	0.14565	0.1458, 0.6452, 0.6192
		60	0.13653	0.1367, 0.7113, 0.7031
		150	**0.12178**	**0.1220, 0.7838, 0.7770**
	Valence	30	0.09890	0.0979, 0.6576, 0.6315
		150	**0.08309**	**0.0823, 0.7789, 0.7715**

Further experiments with the training parameter of the number of epochs were performed. First, the number of epochs was increased by another 30 epochs to a total of 60 epochs. Then, 90 more epochs were added for a total of 150 training epochs. Increasing the number of epochs leads to a longer training period; hence, more learning. Thus, a higher number of epochs can potentially enhance the prediction performance of the CNNs. Table 18 also summarizes the prediction performances after increasing the number of training epochs. The table proves that the higher the number of training epochs, the better the prediction performance. Finally, the MobileNet-v2 CNN

achieved better performance over the ResNet-18 CNN. By increasing the number of epochs to 150, the MobileNet-v2 CNN obtained a testing RMSE, PCC, and CCC of 0.1220, 0.7838, and 0.7770 at arousal predictions, respectively. For valence predictions, it attained a testing RMSE, PCC, and CCC of 0.0823, 0.7789, and 0.7715, respectively. To the best of the author's knowledge, these prediction performances are better than the literature. More performance enhancement is expected if the number of epochs is increased. However, it was decided to stop at 150 epochs due to the time- and power-consuming process of training CNNs. Figure 19 displays a plot of the predicted arousal and valence values against the actual values after 150 training epochs, as per the MobileNet-v2 model, using images of full faces.

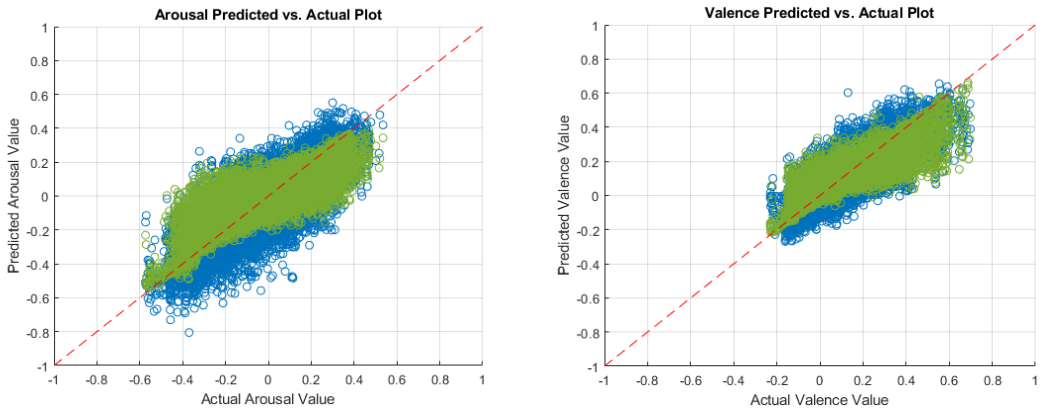


Figure 19. Plots of the predicted versus actual values of (a) arousal and (b) valence predictions by an optimizable ensemble trained on visual features (green), and MobileNet-v2 (150 training epochs) trained on images of full faces (blue).

As described in Section 3.3, decision fusion was applied across the two data modalities (physiological and visual) to fuse the best predictions from the best regressor and CNN. As such, the predictions of the optimizable ensemble regressor for the physiological data were fused with the predictions of the MobileNet-v2 CNN for the visual data, using equation 4 from Section 3.3. Table 19 shows the prediction performances after

multimodal decision fusion. As can be seen, the proposed decision fusion mechanism outperformed the results from the literature.

Table 19. Multimodal Decision Fusion Prediction Performances for Physiological Data and Images of Full Faces

Prediction	Fused Models	RMSE, PCC, CCC
Arousal	Optimizable Ensemble and MobileNet-v2 (150 epochs)	0.0640, 0.9435, 0.9363
Valence		0.0431, 0.9454, 0.9364

Figure 20 displays a plot of the predicted arousal and valence values against the actual values after multimodal decision fusion of the predictions of the optimizable ensemble regressor on physiological data and MobileNet-v2 on images of whole faces. The plot shows a better distribution than the plot from Figure 19. However, the plot from Figure 14, which displays the distribution of the predictions from an optimizable ensemble trained on physiological features, shows a better distribution than Figures 19 and 20. Hence, multimodal decision fusion improved the prediction performance in comparison to the prediction performance of MobileNet-v2.

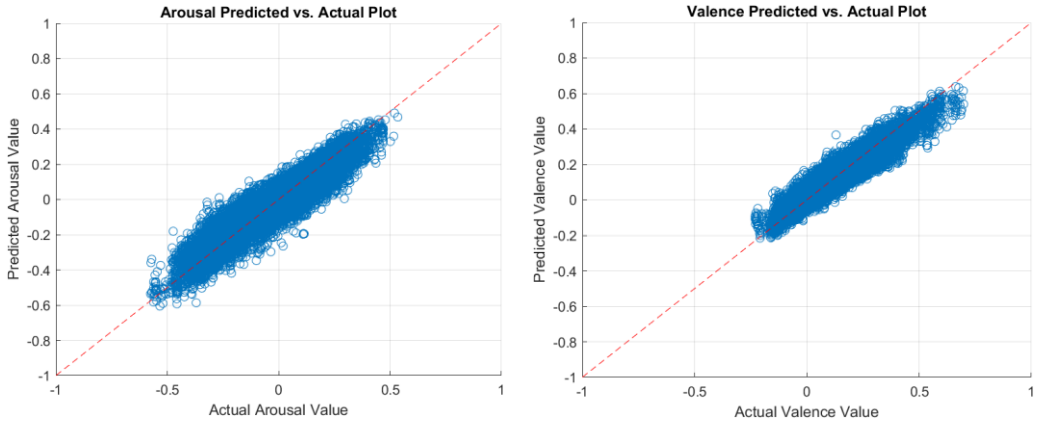


Figure 20. Plots of the predicted versus actual values of (a) arousal and (b) valence after multimodal decision fusion using images of whole faces.

Also, the predictions of the optimizable ensemble regressor for the physiological data were fused with the predictions of the optimizable ensemble regressor for the predesigned visual features, using equation 4 from Section 3.3. Table 20 shows the prediction performances after multimodal decision fusion.

Table 20. Multimodal Decision Fusion Prediction Performances for Physiological and Predesigned Visual Features

Prediction	Fused Models	RMSE, PCC, CCC
Arousal	Optimizable Ensemble (Physiological) and Optimizable Ensemble (Visual)	0.0555, 0.9701, 0.9475
Valence		0.0373, 0.9698, 0.9494

Figure 21 displays a plot of the predicted arousal and valence values against the actual values after multimodal decision fusion using the optimizable ensemble predictions from physiological and predesigned visual features. The plot shows a better distribution than the plot from Figures 19 and 20. The plot from Figure 14 still displays a better distribution than Figures 19-21. Multimodal decision fusion showed an enhanced prediction performance over the optimizable ensemble regressor trained solely on predesigned visual features, as well as MobileNet-v2.

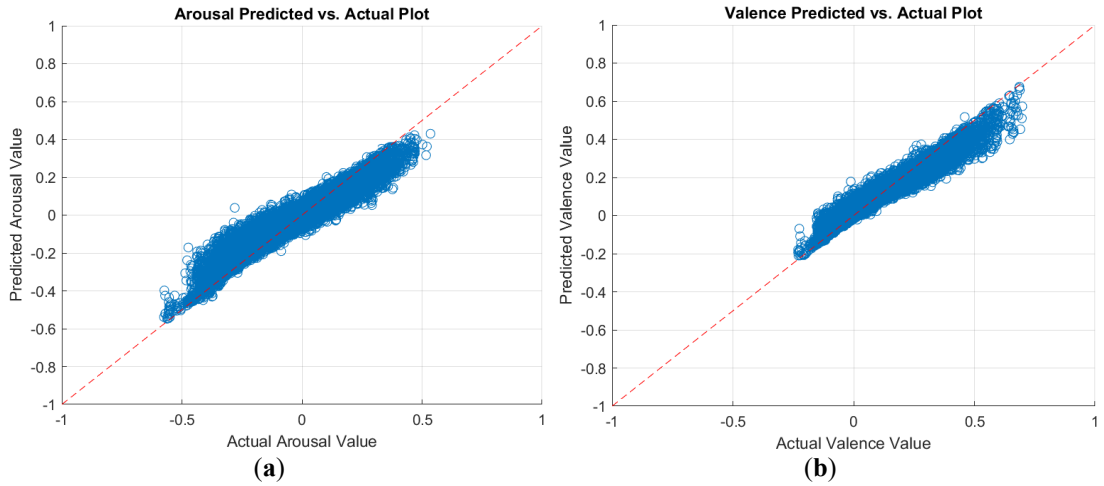


Figure 21. Plots of the predicted versus actual values of (a) arousal and (b) valence after multimodal decision fusion using optimizable ensemble regressors.

5.4. Deep Visual Features

To identify if it is possible to further improve the performance that was reached, this section discusses how deep visual features were extracted from the trained MobileNet-v2. These features have been extracted to replace or to be fused with the predesigned visual features from Section 5.1.2. This work corresponds to stage 10 and 13 of this research and is related to contribution 2 and 3 (see Section 1.3). To the best of the author's knowledge, this combination of predesigned and deep visual features is unique as it was not attempted by other researchers.

After training the MobileNet-v2 CNN on predicting arousal and valence values from images of whole faces, the trained network(s) were used to extract deep visual features from the input images. The deeper layers of the network contain higher-level features that are constructed using the lower-level features from earlier layers. To extract the features of the training and testing images, activations on the global pooling layer, at the end of the network, were used. The global pooling layer pooled features over all spatial locations, providing 1,280 features in total. Table 21 summarizes the dimensions of the extracted sets of deep visual features.

Table 21. Dimensions of Extracted Sets of Deep Visual Features

Data Set	Number of Face Images	Deep Visual Features	Final Dimensions
Original	132751	1280	132751 × 1280
Training	106201	1280	106201 × 1280
Testing	26550	1280	26550 × 1280

Then, the extracted sets of deep visual features were used to train, validate, and test an optimizable ensemble regressor. The optimizable ensemble regressor used the LSBoost algorithm with Bayesian optimization to obtain the best prediction performance. The training process was validated via 5-fold cross validation. In an attempt to improve the prediction performance of the regressor, the same processing techniques as in Sections

4.1.3 and 4.1.4: min-max feature scaling, z-score feature standardization, and outliers-based feature selection were also applied. Table 22 presents the prediction performances for the various scenarios. The rows corresponding to the best performances are displayed in bold font. As can be seen, the aforementioned feature processing steps did not improve the prediction performances of the regressor. The obtained prediction performances are comparable to the prediction performances of MobileNet-v2 (see Table 18).

Table 22. Optimizable Ensemble Prediction Performances on Deep Visual Features

Prediction	Scaled?	Standardized (z-score)?	Feature Selection	Validation RMSE	Testing RMSE, PCC, CCC
Arousal	No	No	None	0.11956	0.1204, 0.7707, 0.7640
	Yes	Yes	None	Lost Data	0.1204, 0.7708, 0.7468
	No	No	5% Outliers (22 Features)	0.16884	0.1690, 0.4472, 0.3316
	No	No	15% Outliers (143 Features)	0.13433	0.1348, 0.7007, 0.6595
Valence	No	No	None	0.081746	0.0812, 0.7761, 0.7530
	Yes	Yes	None	0.082837	0.08237, 0.7685, 0.7429
	No	No	5% Outliers (40 Features)	0.10118	0.0997, 0.6382, 0.5411
	No	No	15% Outliers (381 Features)	0.087476	0.0866, 0.7400, 0.7059

In an attempt to further improve the prediction performance, the deep visual features were further fused with the predesigned visual features. Adding more features as an input to machine learning regressors boosts their performance since this provides more descriptive information about the data. As a result, feature sets of 1,320 features were obtained. Table 23 shows a breakdown of the resulting feature sets of predesigned and deep visual features.

Table 23. Breakdown of Predesigned and Deep Visual Feature Sets

Data Set	Number of Face Images	Predesigned Visual Features	Deep Visual Features	Final Dimensions
Original	132751	40	1280	132751 × 1320
Training	106201	40	1280	106201 × 1320
Testing	26550	40	1280	26550 × 1320

Then, the combined set of features was used to train, validate, and test an optimizable ensemble regressor. The optimizable ensemble regressor used the LSBoost algorithm with Bayesian optimization to obtain the best prediction performance. The training process was validated via 5-fold cross validation. Since the min-max feature scaling, z-score feature standardization, and feature selection steps did not show an improvement in prediction performance earlier (see Table 22), they were not applied after feature fusion. Table 24 presents the prediction performances for feature fusion. The results in the Table 24 show an enhancement in prediction performance over Table 22. In Table 22, the RMSE, PCC, and CCC for arousal were 0.1204, 0.7707, and 0.7640; whereas, for valence, they were 0.0812, 0.7761, and 0.7530.

Table 24. Optimizable Ensemble Prediction Performances on Predesigned and Deep Visual Features

Prediction	Validation RMSE	Testing RMSE, PCC, CCC
Arousal	0.10960	0.1098, 0.8138, 0.7974
Valence	0.079784	0.0784, 0.7947, 0.7834

Figure 22 displays a plot of the predicted arousal and valence values against the actual values using the optimizable ensemble predictions from the deep visual features alone (Table 22) as well as the combined predesigned and deep visual features (Table 24).

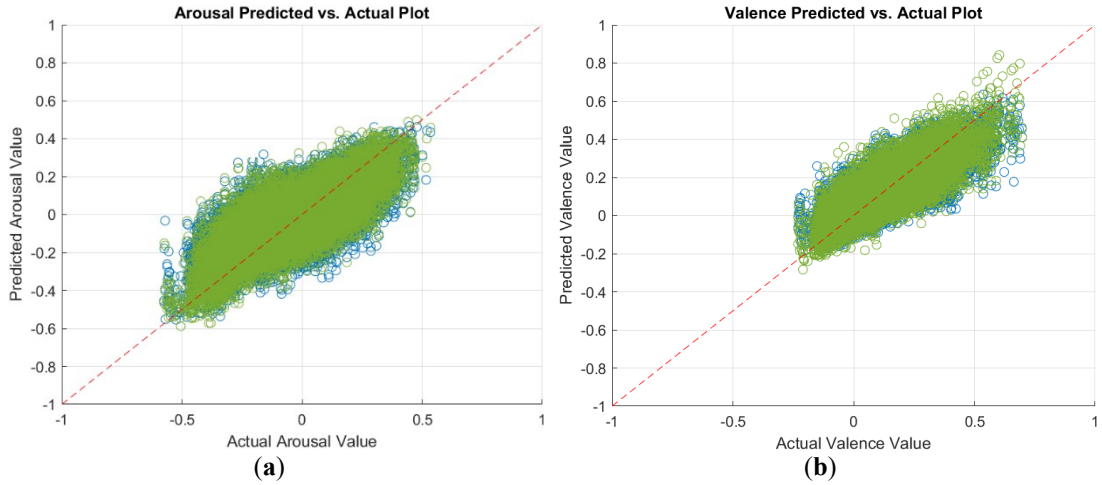


Figure 22. Plots of the predicted versus actual values of (a) arousal and (b) valence by optimizable ensemble regressors using only deep visual features (blue), and a combination of predesigned and deep visual features (green).

In this chapter, visual data recordings from RECOLA were processed. The processed data sets were used to train, validate, and test classical and deep machine learning regressors. Additionally, deep visual features were extracted and used to perform more tests. The acquired prediction performances on visual data were better than the literature. The best prediction performances in the literature, on visual data, were achieved by the authors of [32]. For arousal, they obtained a CCC of 0.699. For valence, they attained a CCC of 0.617. In this study, the reached RMSE was 0.1033, PCC was 0.8498, and CCC was 0.8001 on arousal predictions using predesigned visual features. The achieved RMSE was 0.0702, PCC was 0.8473, and CCC was 0.8053 on valence predictions using predesigned visual features. Applying decision fusion further enhanced the prediction performances (see Table 20).

6. PROOF-OF-CONCEPT FOR AFFECT RECOGNITION USING VISUAL DATA FOR VIRTUAL REALITY APPLICATIONS

This chapter presents stages 2, 6 to 10, and 13 of this research and is related to contributions 1 to 5 (see Section 1.3). It aims to validate the capability of machine learning techniques for AR in an environment similar to the one used in real VR immersion. Note that due to the use of HMD in the envisioned practical application, it is interesting to test the impact of a limited face view (i.e., only the lower part of the face is visible in the presence of an HMD) on the prediction performance. Thus, after processing the video records of RECOLA (see Section 5.1), the extracted images (i.e., video frames) of participants' half faces were inputted into the ResNet-18 and MobileNet-v2 CNNs as a proof-of-concept for predicting arousal and valence values. The trained MobileNet-v2 was then used to extract deep visual features. The predesigned visual features of RECOLA that are associated with the lower half of the face (the only part visible in VR due to the HMD) were also extracted and used as a proof-of-concept for predicting arousal and valence values via classical regression techniques.

6.1. Processing of Visual Data for Virtual Reality (VR) Applications

Given that this study pertains to detecting emotions in users of VR systems who wear HMD that typically cover the eyes and parts (or all) of the nose, the face images from Chapter 5 were further cropped to include only the bottom half of subject faces (i.e., half of the nose, mouth, cheeks, and chin). Figure 23 displays a sample face image before and after face cropping.

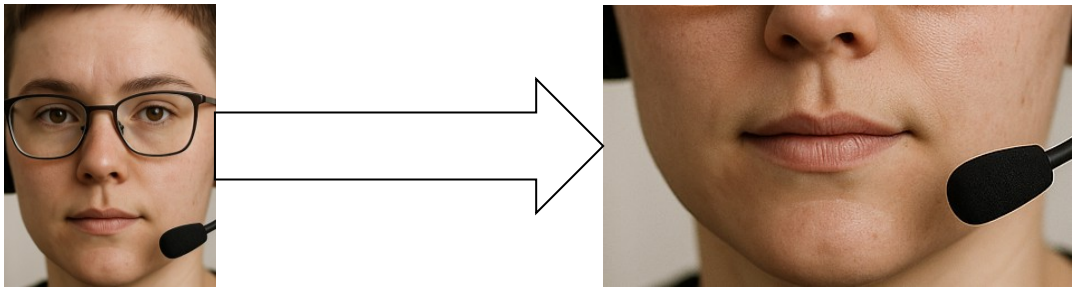


Figure 23. AI-generated face image, mimicking a video frame before and after cropping it to only include the lower half of the face.

The image input layer of the pretrained ResNet-18 and MobileNet-v2 CNNs was replaced with an image input layer of size 140 × 280 × 3 (i.e., RGB layers) to match the size of the processed images/video frames of half faces.

In preparation to perform classical regression, the subset of the predesigned visual features that pertains to the lower half of the face: AU6 (Cheek Raiser), AU11 (Nasolabial Deepener), AU12 (Lip Corner Puller), AU15 (Lip Corner Depressor), AU17 (Chin Raiser), AU20 (Lip Stretcher), AU23 (Lip Tightener), AU24 (Lip Pressor), and AU25 (Lips Part) were extracted from RECOLA. Please note that the other processing steps from Section 5.1 were carried out here as well.

6.2. Regression over Visual Data for Virtual Reality (VR) Applications

This section discusses the deep and classical machine learning results on visual images from the lower half of the face.

6.2.1. Deep Machine Learning Results

Table 25 summarizes the prediction performances for predicting arousal and valence values on images of half faces. MobileNet-v2 appears to slightly outperform ResNet-18. It achieved a testing RMSE, PCC, and CCC of 0.1495, 0.6387, and 0.6081 for arousal; and 0.0996, 0.6453, and 0.6232 for valence. When operating on images of the

lower half of the face, the prediction performance in this study was still better than the ones obtained by other researchers who used complete images of the face.

Table 25. CNN Regression Prediction Performances on Images of Half Faces

Prediction	Network	Training Epochs	Testing RMSE	Testing PCC	Testing CCC
Arousal	MobileNet-v2	30	0.1495	0.6387	0.6081
	ResNet-18		0.1501	0.6323	0.5917
	MobileNet-v2	**150**	**0.1259**	**0.7761**	**0.7717**
	ResNet-18		0.1220	0.7741	0.7622
Valence	MobileNet-v2	30	0.0996	0.6453	0.6232
	ResNet-18		0.1316	0.6403	0.4783
	MobileNet-v2	**150**	**0.0840**	**0.7645**	**0.7510**
	ResNet-18		0.0843	0.7645	0.7480

Figure 24 displays a plot of the predicted arousal and valence values against the actual values, as per the MobileNet-v2 model, using images of half faces. The results obtained using the images of half faces are very similar to those obtained using images of entire faces.

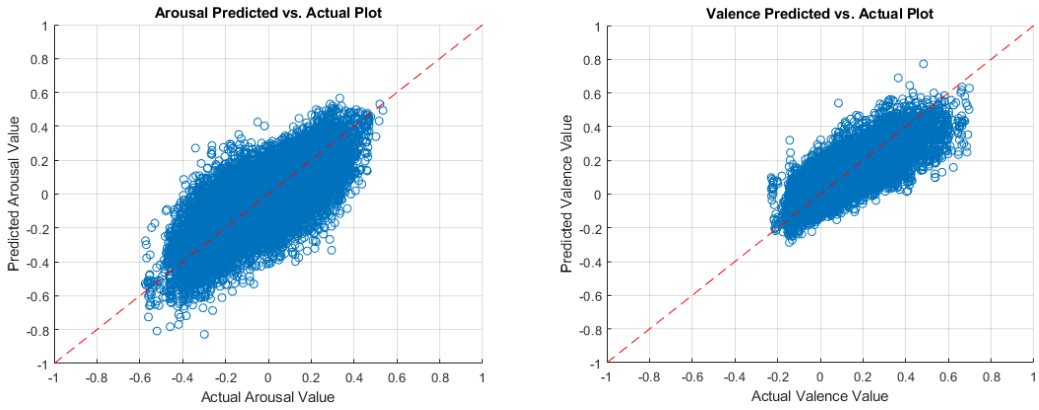


Figure 24. Plots of the predicted versus actual values of (a) arousal and (b) valence as per the MobileNet-v2 model (150 training epochs) using images of half faces.

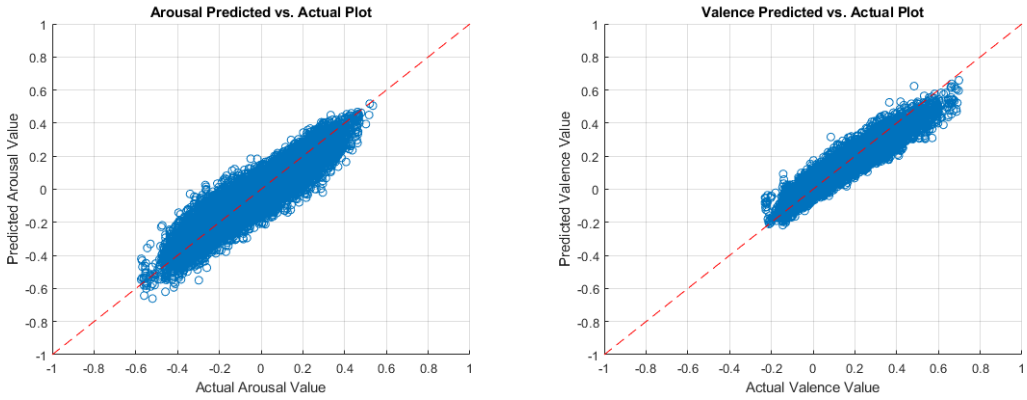
Again, decision fusion was applied to fuse the predictions of the optimizable ensemble regressor that was trained on physiological data with the predictions of the MobileNet-v2 CNN using images of half faces. Table 26 shows the prediction performances after

multimodal decision fusion. As can be seen, the proposed decision fusion mechanism still outperformed the results from the literature, when images of the lower half of the face were used.

Table 26. Multimodal Decision Fusion Prediction Performances using Half Faces

Prediction	Fused Models	RMSE, PCC, CCC
Arousal	Optimizable Ensemble and MobileNet-v2 (half faces, 150 epochs)	0.0658, 0.9391, 0.9339
Valence		0.0441, 0.9450, 0.9319

Figure 25 displays a plot of the predicted arousal and valence values against the actual values after multimodal decision fusion, while using partial facial features of the lower half of the face.



(a) **(b)**

Figure 25. Plots of the predicted versus actual values of (a) arousal and (b) valence after multimodal decision fusion using images of half faces.

6.2.2. Classical Machine Learning Results

Similar to Section 5.4, deep visual features were extracted using the MobileNet-v2 CNN that was trained for predicting arousal and valence values from images of half faces. Then, the predesigned visual features (i.e., facial AUs) that are related to the lower half of the face were combined with the extracted deep visual features. As such, feature sets of 1,289 features (9 AUs and 1,280 deep features) were obtained. The

obtained sets of predesigned and deep visual features were then used to train, validate, and test an optimizable ensemble regressor on the prediction of arousal and valence values. The optimizable ensemble regressor used the LSBoost algorithm with Bayesian optimization to obtain the best prediction performance. The training process was validated via 5-fold cross validation. Table 27 presents the prediction performances for deep visual features extracted from images of half faces. The rows corresponding to the best performances are displayed in bold font. The obtained prediction performances show a slight improvement in RMSE over the prediction performances of MobileNet-v2 (see Table 25). In Table 25, the RMSE, PCC, and CCC for arousal were 0.1259, 0.7761, and 0.7717; whereas, for valence, the values were 0.0840, 0.7645, and 0.7510.

Figure 26 displays a plot of the predicted arousal and valence values against the actual values using the optimizable ensemble predictions from the combined predesigned and deep visual features of images of the lower half of the face.

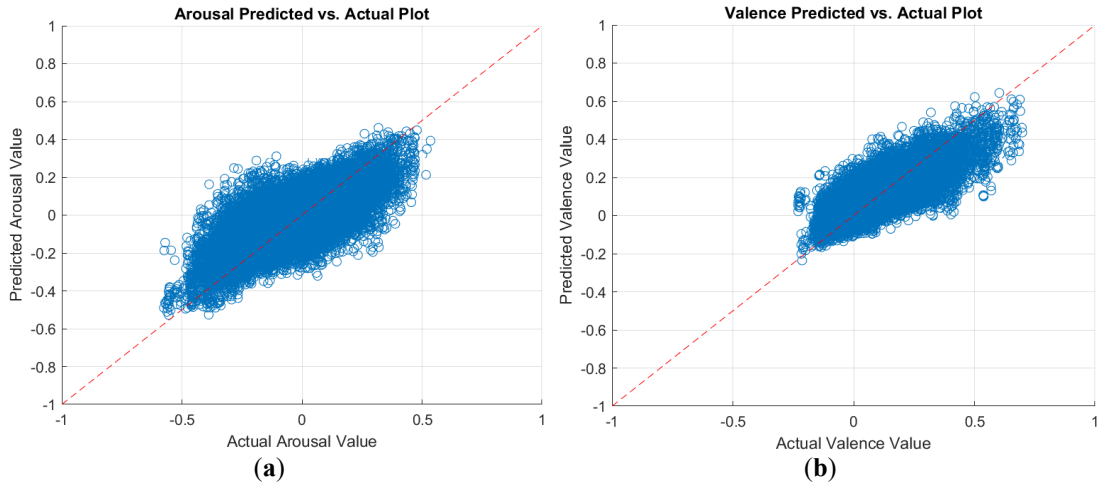


Figure 26. Plots of the predicted versus actual values of (a) arousal and (b) valence by an optimizable ensemble regressor using a combination of predesigned and deep visual features of half-face images.

Table 27. Optimizable Ensemble Prediction Performances on Predesigned and Deep Visual Features of Half Faces

Prediction	Validation RMSE	Testing RMSE, PCC, CCC
Arousal	0.118280	0.1187, 0.7780, 0.7505
Valence	0.083696	0.0832, 0.7633, 0.7360

The best results that were obtained using images of half faces were obtained by the MobileNet-v2 CNN (see Table 25). For arousal, it attained an RMSE, PCC, and CCC of 0.1259, 0.7761, and 0.7717. For valence, it reached an RMSE, PCC, and CCC of 0.0840, 0.7645, and 0.7510.

As a summary of the work performed on RECOLA in this thesis, Table 28 summarizes the approaches that led to the best prediction performances for the physiological and visual data modalities. Table 28 shows that the optimizable ensemble method performed the best on physiological data after feature scaling, z-score standardization, and removal of NaNs. For arousal, a feature selection mechanism allowing only 5% of outliers helped to achieve an RMSE, PCC, and CCC of 0.0168, 0.9965, and 0.9959. For valence, a feature selection mechanism allowing up to 15% of outliers helped to achieve an RMSE, PCC, and CCC of 0.0083, 0.9985, 0.9978. For visual data, the optimizable ensemble method achieved the best performance using predesigned visual features of full faces. It obtained an RMSE, PCC, and CCC of 0.1033, 0.8498, and 0.8001 for arousal predictions; and of 0.0702, 0.8473, and 0.8053 for valence, respectively. The MobileNet-v2 CNN attained the best results using images of half faces, on the other hand. It reached an RMSE, PCC, and CCC of 0.1259, 0.7761, and 0.7717 on arousal; and of 0.0840, 0.7645, and 0.7510 on valence. Therefore, the optimizable ensemble method will be used with feature scaling, standardization, and selection, without NaNs on physiological data in the next stages of this research study. Since the interest of this research study is VR applications, MobileNet-v2 will be used on visual data as it outperformed other methods when operating on images of the lower half of the face.

Table 28. Summary of Best Prediction Performances for the Physiological and Visual Data Modalities

Prediction	Data Type	Data	Scaled?	Standardized (z-score)?	NaNs?	Feature Selection	Regression Type	RMSE, PCC, CCC
Arousal	Physiological	EDA+ECG Features	Yes	Yes	No	5% (50 features)	Optimizable Ensemble	0.0168, 0.9965, 0.9959
	Visual	Images of **Half Faces**	N/A	N/A	N/A	N/A	MobileNet-v2	0.1259, 0.7761, 0.7717
		Predesigned Visual Features of **Full-Face** Images	No	No	N/A	No	Optimizable Ensemble	0.1033, 0.8498, 0.8001
	Physiological + Visual	Predictions on EDA+ECG Features and **Half-Face** Images	N/A	N/A	N/A	N/A	Optimizable Ensemble + MobileNet-v2	0.0658, 0.9391, 0.9339
		Predictions on EDA+ECG Features and Predesigned Features (**Full Faces**)	N/A	N/A	N/A	N/A	Optimizable Ensemble	0.0555, 0.9701, 0.9475
Valence	Physiological	EDA+ECG Features	Yes	Yes	No	15% (103 features)	Optimizable Ensemble	0.0083, 0.9985, 0.9978
	Visual	Images of **Half Faces**	N/A	N/A	N/A	N/A	MobileNet-v2	0.0840, 0.7645, 0.7510
		Predesigned Visual Features of **Full-Face** Images	No	No	N/A	No	Optimizable Ensemble	0.0702, 0.8473, 0.8053
	Physiological + Visual	Predictions on EDA+ECG Features and **Half-Face** Images	N/A	N/A	N/A	N/A	Optimizable Ensemble + MobileNet-v2	0.0441, 0.9450, 0.9319
		Predictions on EDA+ECG Features and Predesigned Features (**Full Faces**)	N/A	N/A	N/A	N/A	Optimizable Ensemble	0.0373, 0.9698, 0.9494

Fusing the predictions of the abovementioned best regressors (i.e., decision fusion), had also been proven to output good prediction performances. The prediction performances achieved after decision fusion were certainly better than those from using visual data alone, as well as the ones reported in the literature by other researchers (see Section 2.3.3) to the best of the author's knowledge. When fusing the best predictions obtained from physiological data and predesigned features of full faces, the obtained RMSE, PCC, and CCC were 0.0555, 0.9701, and 0.9475 for arousal; and 0.0373, 0.9698, and 0.9494 for valence. When combining the predictions obtained from physiological data and images of half faces, the attained RMSE, PCC, and CCC were

0.0658, 0.9391, and 0.9339 on arousal; and 0.0441, 0.9450, and 0.9319 on valence; respectively.

7. PROOF-OF-CONCEPT FOR AFFECT RECOGNITION USING ACOUSTIC DATA

This chapter presents stages 2, 5, 11, and 13 of this research and is related to contributions 1, 4, and 5 (see Section 1.3). As in the case of physiological and visual data, the features from 18 RECOLA audio recordings were considered. All acoustic features were preprocessed by applying time delay and sequencing, feature standardization/normalization, feature scaling, feature selection, annotation labelling, and data shuffling and splitting as detailed in the following sections. After processing the acoustic features, classical regressors were used for predicting arousal and valence values.

7.1. Extraction and Processing of Acoustic Features

The audio recordings of RECOLA were approximately 5 minutes long each, where acoustic features were calculated every 40 milliseconds of recording. That provided a total of approximately 7,500 (5 minutes/40 milliseconds = 7,500) feature vectors per audio recording. The acoustic feature vectors consisted of the low-level descriptors (LLDs) of the Computational Paralinguistics ChallengE (ComParE) [38]. The LLDs contained 4 energy LLDs, 55 spectral LLDs, 6 voicing LLDs, and their first order derivatives for a total of 130 features per feature vector. For more information about RECOLA's acoustic features, please refer to [38]. Table 29 displays the dimensions of the acoustic feature sets that were used.

Table 29. Breakdown of the Acoustic Feature Sets

Feature Set	Number of Feature Vectors	Acoustic Features	Final Dimensions
Original	132,751	130	132,751 × 130
Training	106,201	130	106,201 × 130
Testing	26,550	130	26,550 × 130

To study the impact on the regression performance, multiple preprocessing methods were experimented with. Specifically, standardizing/normalizing the acoustic feature vectors using the z-score method (see equation 5) was explored. This approach standardized the data in each feature vector to have a mean of 0 and an STD of 1.

Furthermore, feature scaling was explored to normalize the range of all features between 0 and 1, for them to equally contribute to the prediction of the regressors. To bring all values into the range [0 1], feature scaling was applied on the acoustic features via the min-max normalization approach (see equation 6). The range of all features should be normalized for them to contribute proportionately to the final estimate of the regressors.

Similar to Section 4.1.4, an outliers-based feature selection mechanism was applied on the acoustic features. When the acceptable percentage of outliers was set to 5%, 28 acoustic features were selected as a result. When the acceptable percentage of outliers was set to 15%, 61 acoustic features were selected as a result.

As proceeded for other data modalities, the first 50 feature vectors were skipped. The remaining annotations were accordingly used to label the corresponding acoustic feature vectors. All labelling and fusion of feature vectors were carried out based on the recordings time. The acoustic feature vectors were shuffled and split, where 80% were used for training and validation, and 20% were used for testing.

7.2. Regression and Results on Acoustic Data

Various tree and ensemble models were trained, validated, and tested for the prediction of arousal and valence values using acoustic features of RECOLA. Specifically, the explored models included fine, medium, coarse, and optimizable tree regressors, as well as ensembled regressors such as boosted trees, bagged trees, and an optimizable ensemble based on bagging and Bayesian optimization. 5-fold cross-validation was performed during training to avoid overfitting. Table 30 and Table 31 show the prediction performances that were reached using the different test scenarios that were ran on acoustic features. The row representing the best performing model is displayed in bold font for each test scenario. The best prediction performance was obtained using an optimizable ensemble regressor trained on normalized and scaled acoustic features. For arousal, it obtained a testing RMSE, PCC, and CCC of 0.1540, 0.5831, and 0.4764. For valence, it attained a testing RMSE, PCC, and CCC of 0.1135, 0.4812, and 0.3237. Unfortunately, a better prediction performance than what is reported in the literature was not achieved.

Table 30. Prediction Performances on Acoustic Features before Feature Processing

Prediction	Standardized (z-score)?	Scaled?	Feature Selection	Regression Model	Validation RMSE	Testing RMSE, PCC, CCC
Arousal	No	No	None	Fine Tree	0.21092	0.2111, 0.3379, 0.3372
	No	No	None	Medium Tree	0.19099	0.1899, 0.3972, 0.3883
	No	No	None	Coarse Tree	0.17325	0.1732, 0.4519, 0.4164
	No	No	None	Optimizable Tree	0.16475	0.1650, 0.4897, 0.4087
	No	No	None	Boosted Trees	0.16506	0.1653, 0.4901, 0.3456
	No	No	None	Bagged Trees	0.15647	0.1561, 0.5636, 0.4693
	No	**No**	**None**	**Optimizable Ensemble**	**0.15450**	**0.1540, 0.5849, 0.4705**
Valence	No	No	None	Fine Tree	0.15617	0.1543, 0.2269, 0.2262
	No	No	None	Medium Tree	0.14112	0.1393, 0.2708, 0.2614
	No	No	None	Coarse Tree	0.12856	0.1269, 0.3162, 0.2757
	No	No	None	Optimizable Tree	0.12213	0.1209, 0.3480, 0.2398
	No	No	None	Boosted Trees	0.15617	0.1543, 0.2269, 0.2262
	No	No	None	Bagged Trees	0.14112	0.1393, 0.2708, 0.2614
	No	**No**	**None**	**Optimizable Ensemble**	**0.12856**	**0.1269, 0.3162, 0.2757**

Table 31. Prediction Performances on Acoustic Features after Feature Processing

Prediction	Standardized (z-score)?	Scaled?	Feature Selection	Regression Model	Validation RMSE	Testing RMSE, PCC, CCC
Arousal	Yes	Yes	None	Fine Tree	0.21093	0.2111, 0.3378, 0.3371
	Yes	Yes	None	Medium Tree	0.19099	0.1899, 0.3972, 0.3883
	Yes	Yes	None	Coarse Tree	0.17325	0.1732, 0.4519, 0.4164
	Yes	Yes	None	Optimizable Tree	0.16480	0.1650, 0.4888, 0.4063
	Yes	Yes	None	Boosted Trees	0.16506	0.1653, 0.4901, 0.3456
	Yes	Yes	None	Bagged Trees	0.15641	0.1559, 0.5658, 0.4719
	Yes	**Yes**	**None**	**Optimizable Ensemble**	**0.15443**	**0.1540, 0.5831, 0.4764**
	Yes	Yes	5%	Fine Tree	0.21884	0.2195, 0.2608, 0.2595
	Yes	Yes	5%	Medium Tree	0.19645	0.1971, 0.3088, 0.2951
	Yes	Yes	5%	Coarse Tree	0.18008	0.1797, 0.3742, 0.3266
	Yes	Yes	5%	Optimizable Tree	0.17279	0.1728, 0.4064, 0.3032
	Yes	Yes	5%	Boosted Trees	0.17321	0.1736, 0.4012, 0.2366
	Yes	Yes	5%	Bagged Trees	0.16665	0.1672, 0.4661, 0.3622
	Yes	**Yes**	**5%**	**Optimizable Ensemble**	**0.16521**	**0.1654, 0.4853, 0.3561**
	Yes	Yes	15%	Fine Tree	0.21510	0.2147, 0.3054, 0.3045
	Yes	Yes	15%	Medium Tree	0.19461	0.1941, 0.3553, 0.3448
	Yes	Yes	15%	Coarse Tree	0.17649	0.1760, 0.4197, 0.3784
	Yes	Yes	15%	Optimizable Tree	0.16852	0.1684, 0.4552, 0.3630
	Yes	Yes	15%	Boosted Trees	0.16749	0.1677, 0.4665, 0.3178
	Yes	Yes	15%	Bagged Trees	0.16078	0.1609, 0.5247, 0.4239
	Yes	**Yes**	**15%**	**Optimizable Ensemble**	**0.15871**	**0.1585, 0.5471, 0.4311**
Valence	Yes	Yes	None	Fine Tree	0.15644	0.1542, 0.2270, 0.2263
	Yes	Yes	None	Medium Tree	0.14126	0.1393, 0.2708, 0.2614
	Yes	Yes	None	Coarse Tree	0.12849	0.1269, 0.3162, 0.2757
	Yes	Yes	None	Optimizable Tree	0.12225	0.1209, 0.3476, 0.2404
	Yes	Yes	None	Boosted Trees	0.12209	0.1211, 0.3566, 0.1669
	Yes	Yes	None	Bagged Trees	0.11589	0.1146, 0.4575, 0.3276
	Yes	**Yes**	**None**	**Optimizable Ensemble**	**0.11509**	**0.1135, 0.4812, 0.3237**
	Yes	Yes	5%	Fine Tree	0.15870	0.1579, 0.1573, 0.1561
	Yes	Yes	5%	Medium Tree	0.14262	0.1417, 0.1940, 0.1816
	Yes	Yes	5%	Coarse Tree	0.13090	0.1301, 0.2316, 0.1867
	Yes	Yes	5%	Optimizable Tree	0.12540	0.1242, 0.2690, 0.1572
	Yes	Yes	5%	Boosted Trees	0.12511	0.1241, 0.2790, 0.1079
	Yes	Yes	5%	Bagged Trees	0.12164	0.1201, 0.3610, 0.2345
	Yes	**Yes**	**5%**	**Optimizable Ensemble**	**0.12018**	**0.1185, 0.3935, 0.2456**
	Yes	Yes	15%	Fine Tree	0.15969	0.1582, 0.1785, 0.1778
	Yes	Yes	15%	Medium Tree	0.14396	0.1427, 0.2102, 0.2004
	Yes	Yes	15%	Coarse Tree	0.13133	0.1297, 0.2546, 0.2124
	Yes	Yes	15%	Optimizable Tree	0.12459	0.1233, 0.2916, 0.1794
	Yes	Yes	15%	Boosted Trees	0.12419	0.1230, 0.3086, 0.1331
	Yes	Yes	15%	Bagged Trees	0.12011	0.1186, 0.3886, 0.2543
	Yes	**Yes**	**15%**	**Optimizable Ensemble**	**0.11879**	**0.1170, 0.4258, 0.2597**

The potential of applying decision fusion using the predictions of the different models trained on visual and acoustic data was further investigated. Since the interest of this study is VR applications, the predictions obtained using the data of half faces were used. Table 32 shows the prediction performances that were reached after decision

fusion. As shown in the table, decision fusion helped to obtain better prediction performances than those that were obtained while using acoustic features alone.

Table 32. Visual and Acoustic Decision Fusion Prediction Performances

Prediction	Data	Fused Models	Testing RMSE, PCC, CCC
Arousal	Deep + Predesigned Visual Features (Half Faces), and Acoustic Features	Optimizable Ensemble and Optimizable Ensemble	0.1248, 0.7771, 0.6725
	Images of Half Faces, and Acoustic Features	MobileNet-v2 and Optimizable Ensemble	0.1195, 0.7849, 0.7246
Valence	Deep + Predesigned Visual Features (Half Faces), and Acoustic Features	Optimizable Ensemble and Optimizable Ensemble	0.0907, 0.7469, 0.6074
	Images of Half Faces, and Acoustic Features	MobileNet-v2 and Optimizable Ensemble	0.0888, 0.7523, 0.6351

The prediction performances on acoustic data were not better than those available in the literature. The best prediction performances in the literature, on acoustic data, were achieved by the authors of [29]. For arousal, they reached an RMSE of 0.107, a PCC of 0.846, and a CCC of 0.846. For valence, they achieved an RMSE of 0.132, a PCC of 0.0.456, and a CCC of 0.450. On the other hand, the obtained RMSE was 0.1540, PCC was 0.5831, and CCC was 0.4764 on arousal predictions using processed acoustic features. The attained RMSE was 0.1135, PCC was 0.4812, and CCC was 0.3237 on valence predictions using processed acoustic features. Fusing the predictions of regressors trained on visual and acoustic features improved the prediction performances (see Table 32).

8. PROOF-OF-CONCEPT FOR AFFECT RECOGNITION USING COMBINED MULTIMODAL FEATURES

This chapter presents stages 2, 5, 12, and 13 of this research and is related to contributions 1, 3, and 6 (see Section 1.3). The processed feature sets that led to the best results from the previous chapters were combined to accomplish multimodal AR through feature-level fusion. After fusing the physiological, visual, and acoustic features, classical regressors were used for predicting arousal and valence values.

8.1. Multimodal Feature Fusion

Similar to stage 1 of research, a novel combination of processing steps was applied to enable the synchronous use of different data modalities (i.e., physiological, visual, and acoustic). This was implemented through time delay and sequencing (see Sections 4.1.1 and 5.1.1). The synchronization of data modalities has made it easy to combine multimodal data via early, feature-level or late, decision-level fusion. The processed feature sets that led to the best results from the previous chapters were combined by feature-level fusion to achieve multimodal affect recognition. As such, the feature sets of 50 physiological features (5% outliers-based feature selection), 40 predesigned full-face visual features, and 130 acoustic features were combined for arousal predictions. The feature sets of 103 physiological features (15% outliers-based feature selection), 40 predesigned full-face visual features, and 130 acoustic features were fused for valence predictions. For half faces, the 9 predesigned visual features that pertain to the lower half of the face (see Section 6.1) were used.

8.2. Regression over Multimodal Data

For the prediction of arousal and valence values from the multimodal feature sets, an optimizable ensemble regressor based on bagging and Bayesian optimization was used. Also, the same tree regressors and ensemble regression models, as in previous sections (4.2.1, 5.2.1, 6.2.2, and 7.2), were experimented with, for comparison purposes. 5-fold cross-validation was implemented during training to avoid overfitting.

8.3. Results on Multimodal Features

In this section, the regression models were trained on the combined feature sets of RECOLA's physiological, visual, and acoustic features. This work corresponds to stages 5, 12, and 13 of this research and is related to contribution 1 and 6 (see Section 1.3). As described in Section 8.2, nine classical regressors: fine tree, medium tree, coarse tree, optimizable tree, boosted trees, bagged trees, and optimizable ensemble were tested for predicting arousal and valence values from combined features. This work was repeated using the visual features of full faces, and again, using the visual features of half faces. Table 33 summarizes the validation and testing prediction performances of these regression models for full-face visual features. When using full-face visual features, the attained testing RMSE, PCC, and CCC were 0.0412, 0.9806, and 0.9733 on arousal predictions, respectively. The obtained testing RMSE, PCC, and CCC were 0.0256, 0.9821, and 0.9784 on valence predictions, respectively. These performances were obtained using an optimizable ensemble regressor with bagging and Bayesian optimization.

Table 33. Multimodal Feature Fusion Prediction Performances with Full-Face Features

Prediction	Physiological Features	Visual Features	Acoustic Features	Features	Regression Type	Validation RMSE	Testing RMSE, PCC, CCC
Arousal	50 processed features, with 5% feature selection	40 predesigned full-face features	130 processed features	220	Fine Tree	0.080246	0.0751, 0.9201, 0.9200
	50 processed features, with 5% feature selection	40 predesigned full-face features	130 processed features	220	Medium Tree	0.085255	0.0806, 0.9061, 0.9054
	50 processed features, with 5% feature selection	40 predesigned full-face features	130 processed features	220	Coarse Tree	0.099865	0.0956, 0.8634, 0.8588
	50 processed features, with 5% feature selection	40 predesigned full-face features	130 processed features	220	Optimizable Tree	0.079706	0.0744, 0.9218, 0.9217
	50 processed features, with 5% feature selection	40 predesigned full-face features	130 processed features	220	Boosted Trees	0.13941	0.1396, 0.7026, 0.5643
	50 processed features, with 5% feature selection	40 predesigned full-face features	130 processed features	220	Bagged Trees	0.054582	0.0488, 0.9700, 0.9625
	50 processed features, with 5% feature selection	**40 predesigned full-face features**	**130 processed features**	**220**	**Optimizable Ensemble**	**0.047273**	**0.0412, 0.9806, 0.9733**
Valence	103 processed features, with 15% feature selection	40 predesigned full-face features	130 processed features	273	Fine Tree	0.049048	0.0449, 0.9385, 0.9385
	103 processed features, with 15% feature selection	40 predesigned full-face features	130 processed features	273	Medium Tree	0.052367	0.0487, 0.9267, 0.9262
	103 processed features, with 15% feature selection	40 predesigned full-face features	130 processed features	273	Coarse Tree	0.064349	0.0596, 0.8868, 0.8837
	103 processed features, with 15% feature selection	40 predesigned full-face features	130 processed features	273	Optimizable Tree	0.048746	0.0440, 0.9413, 0.9413

Prediction	Physiological Features	Visual Features	Acoustic Features	Features	Regression Type	Validation RMSE	Testing RMSE, PCC, CCC
	103 processed features, with 15% feature selection	40 predesigned full-face features	130 processed features	273	Boosted Trees	0.097509	0.0969, 0.6987, 0.5355
	103 processed features, with 15% feature selection	40 predesigned full-face features	130 processed features	273	Bagged Trees	0.030215	0.0268, 0.9812, 0.9761
	103 processed features, with 15% feature selection	**40 predesigned full-face features**	**130 processed features**	**273**	**Optimizable Ensemble**	**0.028955**	**0.0256, 0.9821, 0.9784**

Table 34 summarizes the validation and testing prediction performances of the abovementioned regression models for half-face visual features. When using half-face visual features, the attained testing RMSE, PCC, and CCC were 0.0373, 0.9835, and 0.9786 on arousal predictions, respectively. The reached testing RMSE, PCC, and CCC were 0.0179, 0.9928, and 0.9896 on valence predictions, respectively. These performances were also obtained using an optimizable ensemble regressor with bagging and Bayesian optimization. These performances are better than those from the literature [23,26-36], who performed more complex processing and feature extraction. The best prediction performances, from the literature, on multimodal data were obtained by the authors of [34] for arousal, and [29] for valence predictions. The authors of [34] achieved a CCC of 0.833 on arousal predictions. The authors of [29] attained an RMSE of 0.100, a PCC of 0.689, and a CCC of 0.687 on valence predictions.

Figure 27 displays a plot of the predicted arousal and valence values against the actual values using the optimizable ensemble predictions from the combined physiological, acoustic, and visual features of full faces (Table 33) as well as half faces (Table 34).

Table 34. Multimodal Feature Fusion Prediction Performances with Half-Face Features

Prediction	Physiological Features	Visual Features	Acoustic Features	Features	Regression Type	Validation RMSE	Testing RMSE, PCC, CCC
Arousal	50 processed features, with 5% feature selection	9 predesigned half-face features	130 processed features	189	Fine Tree	0.072235	0.0627, 0.9443, 0.9443
	50 processed features, with 5% feature selection	9 predesigned half-face features	130 processed features	189	Medium Tree	0.079049	0.0704, 0.9287, 0.9282
	50 processed features, with 5% feature selection	9 predesigned half-face features	130 processed features	189	Coarse Tree	0.095703	0.0878, 0.8860, 0.8826
	50 processed features, with 5% feature selection	9 predesigned half-face features	130 processed features	189	Optimizable Tree	0.071789	0.0620, 0.9456, 0.9456
	50 processed features, with 5% feature selection	9 predesigned half-face features	130 processed features	189	Boosted Trees	0.14131	0.1414, 0.6934, 0.5486
	50 processed features, with 5% feature selection	9 predesigned half-face features	130 processed features	189	Bagged Trees	0.050657	0.0439, 0.9755, 0.9701
	50 processed features, with 5% feature selection	**9 predesigned half-face features**	**130 processed features**	**189**	**Optimizable Ensemble**	**0.043880**	**0.0373, 0.9835, 0.9786**
Valence	103 processed features, with 15% feature selection	9 predesigned half-face features	130 processed features	242	Fine Tree	0.049249	0.0477, 0.9304, 0.9303
	103 processed features, with 15% feature selection	9 predesigned half-face features	130 processed features	242	Medium Tree	0.052745	0.0513, 0.9185, 0.9179
	103 processed features, with 15% feature selection	9 predesigned half-face features	130 processed features	242	Coarse Tree	0.064735	0.0618, 0.8779, 0.8742
	103 processed features, with 15% feature selection	9 predesigned half-face features	130 processed features	242	Optimizable Tree	0.048834	0.0469, 0.9331, 0.9330
	103 processed features, with 15% feature selection	9 predesigned half-face features	130 processed features	242	Boosted Trees	0.098010	0.0969, 0.6990, 0.5353
	103 processed features, with 15% feature selection	9 predesigned half-face features	130 processed features	242	Bagged Trees	0.030254	0.0271, 0.9804, 0.9756
	103 processed features, with 15% feature selection	**9 predesigned half-face features**	**130 processed features**	**242**	**Optimizable Ensemble**	**0.021396**	**0.0179, 0.9928, 0.9896**

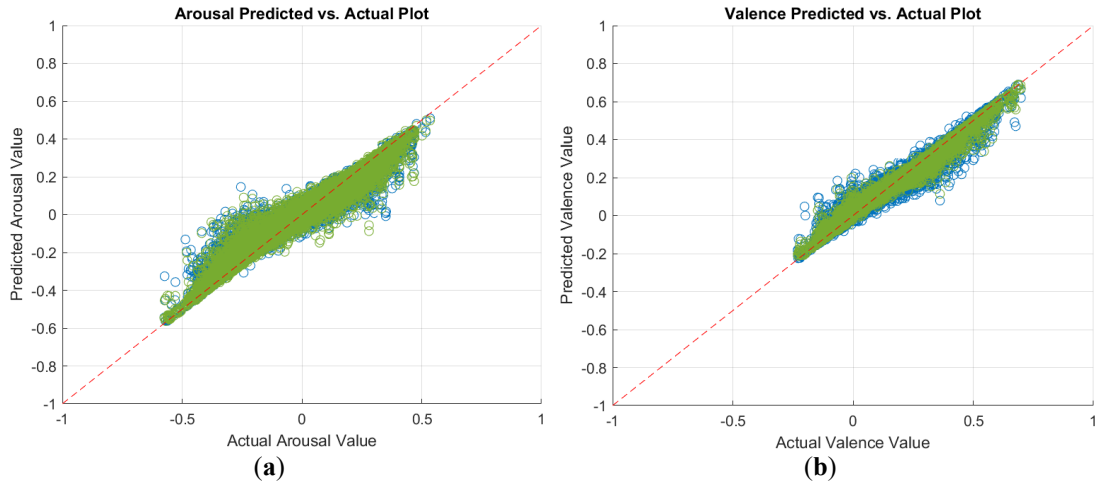


Figure 27. Plots of the predicted versus actual values of (a) arousal and (b) valence by optimizable ensemble regressors using only combined physiological, acoustic, and visual features of full faces (blue) and half faces (green).

In this chapter, feature fusion was applied to combine physiological, visual, and acoustic data from RECOLA in order to perform multimodal AR. The performed tests included combinations of physiological features, full-face or half-face visual features, and acoustic features. For the combination containing half-face visual features, the best prediction performances were reached using an optimizable ensemble regressor with bagging and Bayesian optimization. For arousal, it achieved an RMSE of 0.0373, a PCC of 0.9835, and a CCC of 0.9786. For valence, it attained an RMSE of 0.0179, a PCC of 0.9928, and a CCC of 0.9896.

9. AFFECT RECOGNITION USING DATA COLLECTED IN VIRTUAL REALITY

This chapter presents stages 14 to 19 of this research project and is related to contribution 1, 3, and 5 to 9 (see Section 1.3). This research is part of a broader interdisciplinary project involving the Institute of Mental Health Research at the Royal Hospital, the University of Ottawa's Department of Psychiatry, and the Department of Psychoeducation and Psychology, and the Department of Computer Science and Engineering at UQO. The goal is to create and develop a new adaptable intervention for treating cognitive impairments using VR, leveraging combined approaches from computer science and psychology. To implement this solution in a practical setting, a newly developed VR system for cognitive training for individuals with schizophrenia was chosen.

Schizophrenia is a mental health disorder that is diagnosed based on the presence of delusions, hallucinations, as well as disorganized thoughts, speech, and behavior [66]. Such symptoms severely impact personal, family, social, educational, and occupational aspects in the lives of people with schizophrenia. Schizophrenia affects approximately 0.32% of people worldwide, and at a rate of 0.45% among adults. It is not as common as other mental disorders, but it is one of the most severe, incapacitating, and complex to manage. People with schizophrenia also face constant difficulties with their cognitive or thinking skills including memory, attention, and problem-solving.

VR can help to transfer learned skills of traditional cognitive remediation programs [43]. A research group recently developed VR training scenarios to implement

cognitive remediation strategies in people with schizophrenia [43]. This study aims to improve VR for cognitive interventions for people with schizophrenia by adding AR to modulate the difficulty of the cognitive training tasks used in cognitive remediation. This work acts as a proof-of-concept for predicting affective states from arousal and valence values via physiological and visual data of users immersed in VR experiences. In the future, beyond the scope of this work, the predicted affective states could be used to modulate in real time the cognitive load of users and make the training more adaptive.

9.1. Data Collection in Virtual Reality (VR)

In the proof-of-concept studies, only annotated data from the RECOLA data set, which was collected in a controlled environment, were worked on. This chapter introduces work on predicting affective states through physiological, and visual data collected from VR immersions as part of contribution 8 (see Section 1.3). A research protocol was submitted and approval from REB or IRB was obtained in order to collect real data. Colleagues at the Cyberpsychology Laboratory of UQO collected physiological and visual data in a VR setting, where 10 participants were immersed in three different virtual environments and completed various tasks. This section will further discuss the data collection strategy and the applied procedures.

9.1.1. Virtual Reality (VR) Data and Equipment

For the VR experiment, the VIVE Eye Pro HMD provided by the High-Tech Computer Corporation (HTC) was used. The collected data contained emotional arousal and valence values based on self-reporting answers, pupillometry data, facial markers, and ECG data.

The visual data consisted of video recordings of the lower half of participants' faces, facial markers, and pupillometry recordings. The facial markers included jaw forward

movement, jaw movement to the left/right, jaw openness, mouth shape, mouth pout, mouth movement to the lower right/lower left, mouth smile to the right/left, mouth sad pose to the right/left, right/left cheek puffiness, placement of lower/upper lip inside, overlay of lower/upper lip, cheek suck, movement of the lower right/lower left lip downwards, movement of the upper right/upper left lip upwards, and mouth philtrum movement to the right/left. An eye tracking sensor, embedded into the VR HMD, was used to measure pupillometry features such as eye openness, pupil diameter, and pupil position in the X and Y directions [67]. The eye tracker operates at a frequency of 120 hertz.

The ECG recordings represented physiological data. A T31 coded™ wireless transmitter belt from Polar Electro was used to collect ECG data. Participants were asked to wear the belt to measure the electrical activity of their hearts. The belt was non-invasive and was worn near the chest area of participants' bodies. The ECG data were used to extract estimates of participants' HR and HRV.

9.1.2. Virtual Reality (VR) Experiments

Tasks in the VR prototype included social and cognitive training in three different environments: a restaurant, a bus, and an apartment. The restaurant environment provided verbal memory exercises, whereas the bus environment provided spatial and verbal memory, as well as social cognition exercises. Lastly, the apartment environment provided attention, memory, daily functioning skills, problem solving, and social cognition tasks. These virtual environments were developed by the research team of Dr. Bouchard in the Cyberpsychology Laboratory of UQO. For more information about the VR tasks of each environment, refer to Appendix C.

Each environment had four levels. All 10 participants had to do the first level of each environment in order to hear and understand the proper instructions from the virtual

agents. In total, each participant was involved in eight immersions: four restaurant levels, two apartment levels, and two bus levels. During these immersions, participants were repeatedly asked to stop the task at hand to answer a few questions. Thus, participants were randomly assigned to three different groups, and all approximately had 45 minutes of immersion time as well as 30 pauses to answer the experimenter's questions. Additionally, there was a baseline of 30 seconds in which the participants were immersed in VR, but did not move around or talk. Participants were also asked to walk around the virtual environment for a total of 1 minute and 30 seconds to familiarize themselves to moving around with the VR controllers. For a detailed view of the data collected with the participants, randomly assigned to three groups experiencing the different levels of difficulty of the VR tasks, please refer to Appendix C.

The same four self-reporting questions were asked on 30 occasions (i.e., 30 pauses to answer questions) to determine the affective state of participants. These occasions were designed not to interfere with the cognitive tasks, which include instructions, verbal exchanges as part of cognitive tasks, and/or performing actions (e.g., not asking question at the moment participants are receiving, listening to, or talking with a virtual character). After testing each level of the virtual environments systematically, it was found that the most amount of pauses possible for every group was 30. To reach this number, each level was deconstructed, and every potential logical point to pause was noted. More precisely, approximately 30 seconds between these points were kept, if possible, while avoiding the addition of pauses when the virtual characters were speaking to the participant or when the participant was already talking to the virtual characters. For example, in the restaurant environment, if a participant needed to listen to instructions at the front desk, walk to a table (5-30 seconds depending on the participant), take an order, walk back to the front desk (5-30 seconds), and repeat the order to the cook, there could be a total of three pauses:

1. when they walk to the table,
2. when they walk back to the front desk, and
3. immediately after repeating the order.

The actual virtual environments could not be paused so pauses were not added when the characters were speaking or providing instructions. Every group was compared regarding their total amount of potential pauses in their respective levels to come up with a total of 30 pauses.

Each participant was required to answer the self-reporting questions according to how they felt when they were asked to stop. The four self-reporting questions were:

1. On a scale from 0 to 10, how hard were you trying just before we stop?
2. What emotion were you feeling the most: sadness, calmness, happiness, or anger?
3. Depending on the previous answer, on a scale from 0 to 10, how positive or negative were you feeling?
4. On a scale from 0 to 10, how energized were you feeling?

Participants provided their answers to the above questions verbally, and their answers were recorded by the experimenter. The first question was asked to assess cognitive effort in future studies, while the remaining questions were used in this VR study.

9.1.3. Annotation Labelling of Virtual Reality (VR) Data

After the VR sessions had been completed, the Circumplex model of affective states (see Figure 1) was used to classify the emotions participants were feeling during the task. Within the Circumplex model, affective states are characterized as discrete points in a 2D space of valence and arousal axes. Valence is used to rate the positivity of the

affective state. Arousal is used to rate the activity/energy level of the affective state. There are four quadrants in the Circumplex model: low arousal/low valence (sad), low arousal/high valence (relaxed), high arousal/low valence (angry), and high arousal/high valence (happy). The answers from the second question helped us to determine which quadrant of the Circumplex model did participants' emotions belong. The answers to the third and fourth questions helped us to estimate valence and arousal values, respectively. According to the Circumplex model, the highest arousal/valence value is 1 and the lowest arousal/valence value is -1. A value of 0 translates to neutral emotions. Hence, the valence and arousal values were estimated using the following equations:

$$valence = \pm\frac{1}{10} \times answer\ to\ \boldsymbol{third}\ self-reporting\ question \tag{7}$$

$$arousal = \pm\frac{1}{10} \times answer\ to\ \boldsymbol{fourth}\ self-reporting\ question \tag{8}$$

If the participant answered 'sadness' to the second self-reporting question, then both valence and arousal values were negative. If they answered 'calmness', the resulting valence value was positive and the arousal value was negative. If the answer was 'happiness', then both valence and arousal values were positive. Finally, for 'anger', the valence value was negative, whereas the arousal value was positive.

9.1.4. Comparison of Data Sets

This section provides a comparison to highlight the similarities and differences between the RECOLA data set and the data set that was collected for this thesis as part of a VR study. These details are critical to justify and clarify any steps that have been done differently in the VR study from the proof-of-concept studies.

The RECOLA data set provides continuous physiological, visual, and acoustic recordings and features, with their arousal/valence annotations for 18 participants. These recordings and features were all sampled every 40 milliseconds. After processing and synchronization between the different data modalities, this had supplied a training data set of 106,201 samples and a testing data set of 26,550 samples, totaling in 132,751 samples. The physiological data set from RECOLA included sample vectors of 118 variables (1 EDA reading, 62 EDA features, 1 ECG reading, and 54 ECG features). The visual data set from RECOLA contained feature vectors of 40 facial features each.

The data set for this VR study provided continuous visual features (eye and facial markers) collected during immersive VR experiences of 10 participants. However, the provided physiological features were only calculated at certain points in time. Similarly, the self-reporting questionnaires used to derive the arousal and valence annotations were only asked at specific occasions. The questionnaires were asked around 30 times to each participant, while they were immersed. The physiological/ECG features were later calculated 40 milliseconds before each questionnaires period. The nature of the data collection process has created a big difference in the amount of data samples between the physiological and visual data modalities. Note that the facial features of RECOLA differ from the ones collected for the VR study. Nonetheless, the data collected for the VR study included pupillometry/eye features. Also, the physiological data as well as arousal and valence labels were not continuous as in RECOLA. This has added a requirement for additional processing steps on top of what was applied in the proof-of-concept studies (Chapters 4 to 8). In this case, the final training data set contained 218 samples, and the testing data set contained 55 samples, totaling 273 samples. The physiological data set included sample vectors of 4 variables (ECG features). The visual data set contained feature vectors of 30 facial features (4 pupillometry features and 26 facial features) each.

The next section will explain how the collected VR data were processed in preparation for machine learning regression. Collected data were processed in such a manner to match the format of the data recordings of RECOLA as much as possible. This has made the training and testing processes easier, where the processed VR data recordings were used as input to the same regressors as the ones selected as part of the proof-of-concept studies. As such, an optimizable ensemble regressor was used for the collected ECG data as well as eye/facial features.

9.2. Processing and Cleaning of Virtual Reality (VR) Data

The decision was made not to proceed with the video recordings of the facial expressions since the proof-of-concept studies showed that combinations of physiological and visual features perform better for AR.

In terms of physiological data, four ECG-related measurements were calculated: average HRV at very low-frequency power, average HRV at low-frequency power, average HRV at high-frequency power, and average blood volume pulse/heart rate. These measurements were calculated 40 milliseconds before each self-reporting questionnaire (see Section 9.1.2) to capture the state of participants without being influenced by the potential psychological impact of completing the questionnaires.

In terms of visual data, the collected eye and facial features were worked with. Eye and facial data were collected throughout the entire VR sessions, with no pauses for when questions were being asked. Each sample consists of a feature vector of 30 eye and facial features (see Section 9.1.1). Since visual data were collected the whole time, there was a need to manually clean the data by removing any readings that were collected during the questionnaires as they do not reflect the participants' VR

immersion experience. To produce clean data sets for AR, the visual data were manually processed in the following manner:

- The samples at the beginning of each VR immersion session, which consist of baseline data where the participants were not moving and/or getting familiar with the environment, were removed. These samples served no significance for the purposes of this study as they were not associated with any arousal/valence annotations.
- The samples surrounding (i.e., within the same second in time) and following the pauses (i.e., until the end of each VR immersion session) for self-reporting questionnaires were all removed to avoid readings during the questionnaires period.

Afterwards, feature processing steps such as the removal of NaNs by replacing them with zeros, as well as feature scaling and standardization/normalization via formulas (5) and (6) were performed respectively (see Section 4.1.3). Due to the nature of the data collection process, one sample of physiological data was acquired per VR immersion session, while a sequence of samples of visual data was acquired. Similarly, for each arousal/valence annotation, one sample of physiological measurements, and a sequence of visual (eye and facial) features were obtained. In total, the acquired data set consisted of 273 samples of physiological data, and 273 sequences of visual (eye and facial) data. The length of the visual sequences varied depending on the length of the corresponding VR sessions. The 273 sequences of visual data corresponded to 97,061 eye and facial feature vectors.

The arousal/valence annotations were based on the participants answers to self-reporting questionnaires during pauses in specific VR environments and levels, after completing specific VR tasks (i.e., immersion sessions). The physiological and visual

data recordings of all participants were grouped by those specific sessions and then shuffled to ensure randomization. Hence, the performed AR is dependent on the VR environment, level, and session/task (i.e., immersion sessions). It is independent of participants as the data from multiple participants were grouped and shuffled.

9.3. Virtual Reality (VR) Data Analysis

This section provides an analysis of the physiological and visual features that were collected from the VR experiments.

9.3.1. Features from Virtual Reality (VR) Data

To summarize, the collected features included 4 physiological/ECG features, and 30 visual features (4 pupillometry features and 26 facial features). The four ECG features included three averages of HRV at different frequencies, and averages of HR. Figure 28 shows a scatter plot of the samples of the collected physiological feature vectors. The four pupillometry/eye features were eye openness, pupil diameter, and pupil position in the X and Y directions. The 26 facial features consisted of sequences of the aforementioned facial markers (see Section 9.1.1). Figure 29 shows a plot of a sample sequence of visual (eye and facial) features.

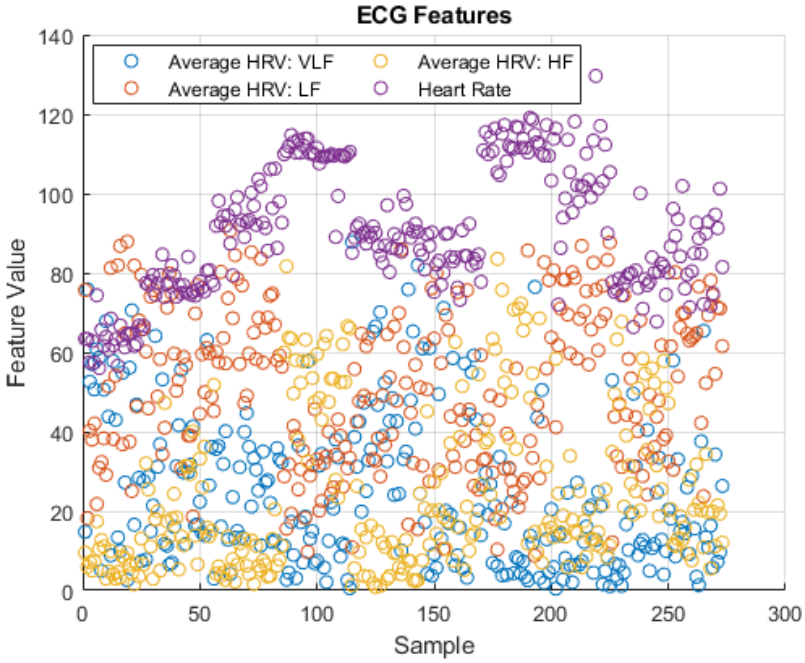


Figure 28. Collected samples of physiological features.

Figure 29. Sample sequence of visual features.

9.3.2. Statistics of Features from Virtual Reality (VR) Data

This section includes an analysis of the statistics of the physiological features, and the averages of visual feature sequences that were collected during VR immersions. The analysis was conducted on the original (i.e., before processing) data sets of 273 samples.

Statistical measures such as the mean, variance, STD, and coefficient of variation (CV) of the collected feature vectors as well as the arousal and valence annotations were calculated to study the amount of variation between the samples of each feature vector. It is important to ensure that there are enough variations between the samples of each feature vector to enable machine learners to generalize adequately. A low STD suggests that the values are generally near the mean of the feature vector, whereas a high STD implies that the values are dispersed over a broader range [68]. The CV is a relative metric used to assess variability by comparing the STD to the mean. Since it is a normalized, unitless measure, it enables comparisons of variability across different data sets or characteristics. To calculate it, one can divide the standard deviation by the mean. A higher CV suggests that the STD is large relative to the mean. A CV of 1 (or 100%) means the STD equals the mean. Values below 1 indicate less variability (STD is smaller than the mean), which is common. Conversely, values above 1 signify greater variability (standard deviation exceeds the mean). In general, a higher CV reflects more relative variability.

Table 35 summarizes the statistical measures of the collected feature vectors and arousal and valence labels. The statistics from this table show that most of the visual features in the collected data set as well as the arousal and valence annotations had a good amount of variation. On the other hand, the statistics show that the physiological feature vectors had little variation between the samples.

Table 35. Statistical Measures of the Collected Feature Vectors

Data Modality	Variable/Feature	Statistical Measure			
		Mean	**Variance**	**STD**	**CV**
Physiological (ECG)	Average HRV: VLF	24.8670	376.5488	19.4049	0.7804
	Average HRV: LF	50.7664	394.5704	19.8638	0.3913
	Average HRV: HF	24.3939	411.8125	20.2932	0.8319
	Average HR	90.2560	256.4339	16.0136	0.1774
Visual (Pupillometry)	Eye Openness	0.8931	0.0215	0.1465	0.1640
	Pupil Diameter	3.4360	0.7845	0.8857	0.2578
	Pupil X Position	0.0899	0.0276	0.1660	1.8465
	Pupil Y Position	-0.0611	0.0248	0.1576	-2.5794
Visual (Facial)	Jaw Forward	0.2146	0.0444	0.2106	0.9814
	Jaw Right	0.0375	0.0010	0.0321	0.8560
	Jaw Left	0.0277	0.0006	0.0252	0.9098
	Jaw Open	0.0222	0.0014	0.0369	1.6622
	Mouth Ape Shape	0.0276	0.0051	0.0716	2.5942
	Mouth O Shape	0.0091	0.0004	0.0190	2.0879
	Mouth Pout	0.0400	0.0088	0.0940	2.3500
	Mouth Lower Right	0.0394	0.0012	0.0345	0.8756
	Mouth Lower Left	0.0305	0.0007	0.0267	0.8754
	Mouth Smile Right	0.0126	0.0030	0.0545	4.3254
	Mouth Smile Left	0.0103	0.0026	0.0505	4.9029
	Mouth Sad Right	0.1452	0.0308	0.1754	1.2080
	Mouth Sad Left	0.1489	0.0328	0.1811	1.2163
	Cheek Puff Right	0.0045	0.0005	0.0229	5.0889
	Cheek Puff Left	0.0078	0.0011	0.0332	4.2564
	Mouth Lower Inside	0.0362	0.0059	0.0771	2.1298
	Mouth Upper Inside	0.0154	0.0026	0.0511	3.3182
	Mouth Lower Overlay	0.0440	0.0092	0.0957	2.1750
	Mouth Upper Overlay	0	0	0	0
	Cheek Suck	0.0174	0.0030	0.0543	3.1207
	Mouth Lower Right Down	0.0320	0.0032	0.0565	1.7656
	Mouth Lower Left Down	0.0326	0.0036	0.0599	1.8374
	Mouth Upper Right Up	0.0732	0.0128	0.1132	1.5465
	Mouth Upper Left Up	0.0743	0.0128	0.1130	1.5209
	Mouth Philtrum Right	0.0279	0.0006	0.0245	0.8781
	Mouth Philtrum Left	0.0231	0.0004	0.0202	0.8745
Affective State	Arousal	-0.2293	0.3274	0.5722	-2.4954
	Valence	0.3927	0.2092	0.4574	1.1648

Table 36 summarizes the correlation coefficients between the ECG features as well as the arousal and valence labels. Values of the correlation coefficient can range from -1 to +1. A value of -1 indicates perfect negative correlation, while a value of +1 indicates perfect positive correlation. A value of 0 indicates no correlation between the features.

The table shows that the ECG features were weakly correlated with the arousal and valence annotations.

Table 36. Correlation Coefficients between the Physiological/ECG Features as well as the Arousal and Valence Annotations

Variable/Feature	Average HRV: VLF	Average HRV: LF	Average HRV: HF	Average HR	Arousal	Valence
Average HRV: VLF	1	-0.4450	-0.4976	-0.3002	-0.0472	-0.0477
Average HRV: LF	-0.4450	1	-0.5354	-0.1070	0.1539	0.0126
Average HRV: HF	-0.4976	-0.5354	1	0.4030	-0.1080	0.0185
Average HR	-0.3002	-0.1070	0.4030	1	0.0901	-0.0056
Arousal	-0.0472	0.1539	-0.1080	0.0901	1	-0.3457
Valence	-0.0477	0.0126	0.0185	-0.0056	-0.3457	1

Self-evaluation/reporting requires a very high degree of awareness from a person. Not all people can confidently state if they were aroused or not. This was another potential factor of bias in the collected data. In RECOLA, experts were hired to rate the arousal/valence values of participants at a continuous pace (i.e., every 40 milliseconds). Since the evaluation process was not continuous as in RECOLA, the collected data set was missing sufficient arousal/valence labels, unlike the data set in the proof-of-concept studies. One can be aroused one minute and then get their attention distracted. At the end, when asked, they might have forgotten that they were aroused. Thus, the missing link between the features and the arousal/valence annotations is probably because there were no continuous arousal/valence labels.

Figure 30 further displays the correlation matrix between the different physiological/ECG features and the arousal and valence labels.

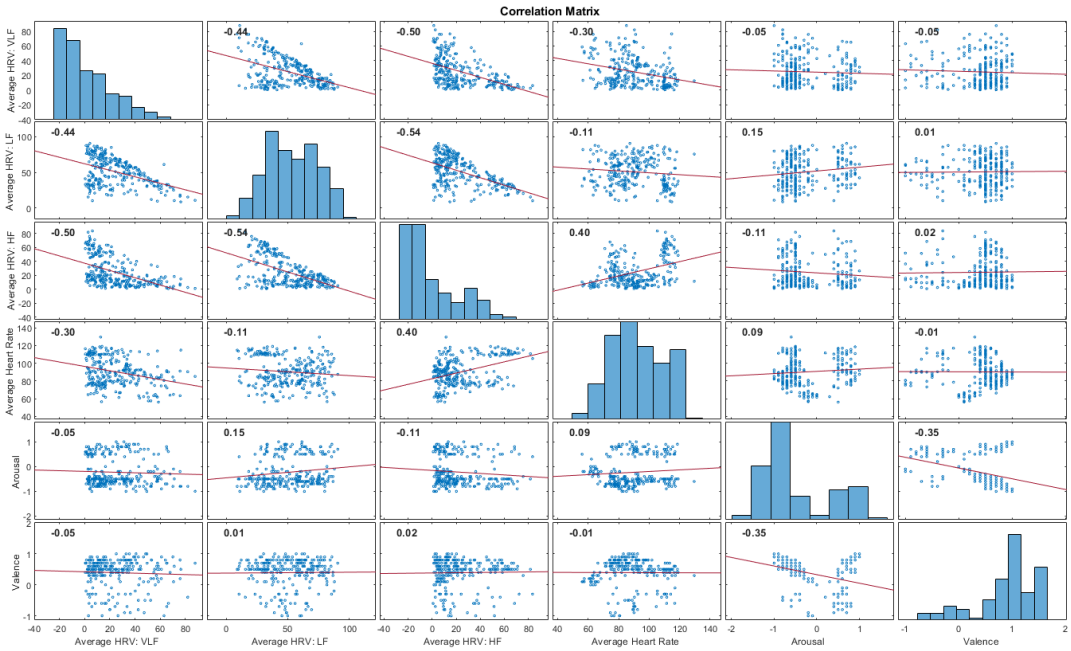


Figure 30. Correlation matrix between the different physiological/ECG features and the arousal and valence annotations.

Table 37 summarizes the correlation coefficients between the pupillometry/eye features as well as the arousal and valence labels. The table shows that the pupillometry/eye features were weakly correlated with the arousal and valence annotations. This is probably due to the lack of sufficient arousal/valence labelling.

Table 37. Correlation Coefficients between the Pupillometry/Eye Features as well as the Arousal and Valence Annotations

Variable/Feature	Eye Openness	Pupil Diameter	Pupil X Position	Pupil Y Position	Arousal	Valence
Eye Openness	1	0.3619	-0.0013	-0.0603	0.0350	0.0373
Pupil Diameter	0.3619	1	0.1780	-0.0255	0.0854	0.0640
Pupil X Position	-0.0013	0.1780	1	0.2203	-0.1129	0.0643
Pupil Y Position	-0.0603	-0.0255	0.2203	1	-0.0451	-0.1250
Arousal	0.0350	0.0854	-0.1129	-0.0451	1	-0.3457
Valence	0.0373	0.0640	0.0643	-0.1250	-0.3457	1

Figure 31 further displays the correlation matrix between the different pupillometry/eye features and the arousal and valence labels.

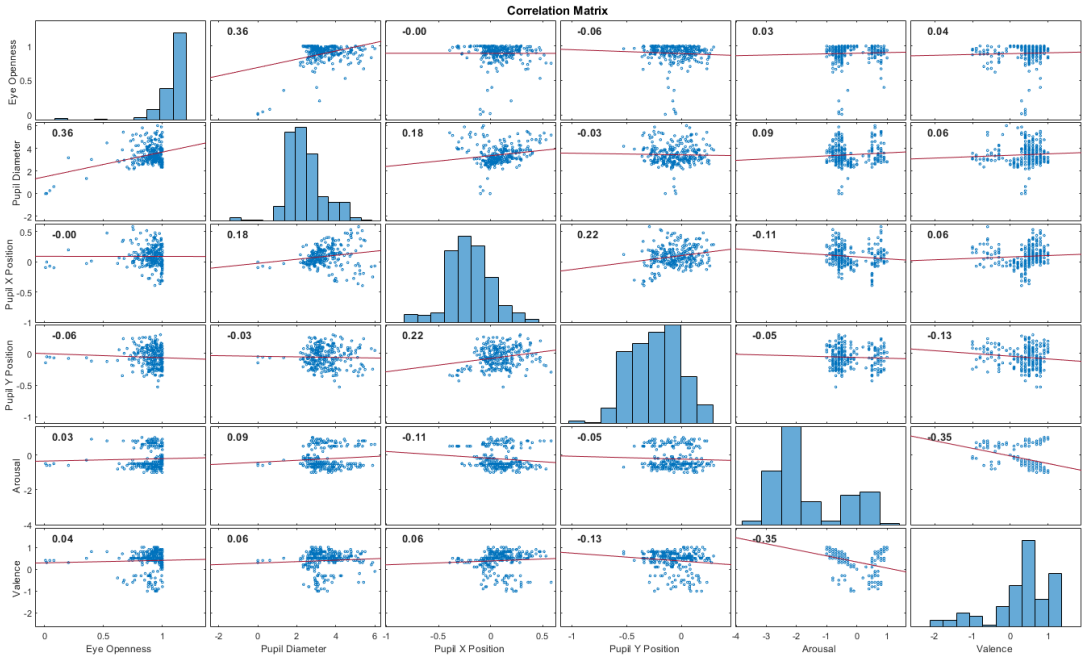


Figure 31. Correlation matrix between the different pupillometry/eye features and the arousal and valence annotations.

Figure 32 displays the correlation matrix between the different facial features and the arousal and valence labels. Note that, in the figure, the Mouth Upper Overlay variable/feature has been removed as it only contained zeros and has no significance.

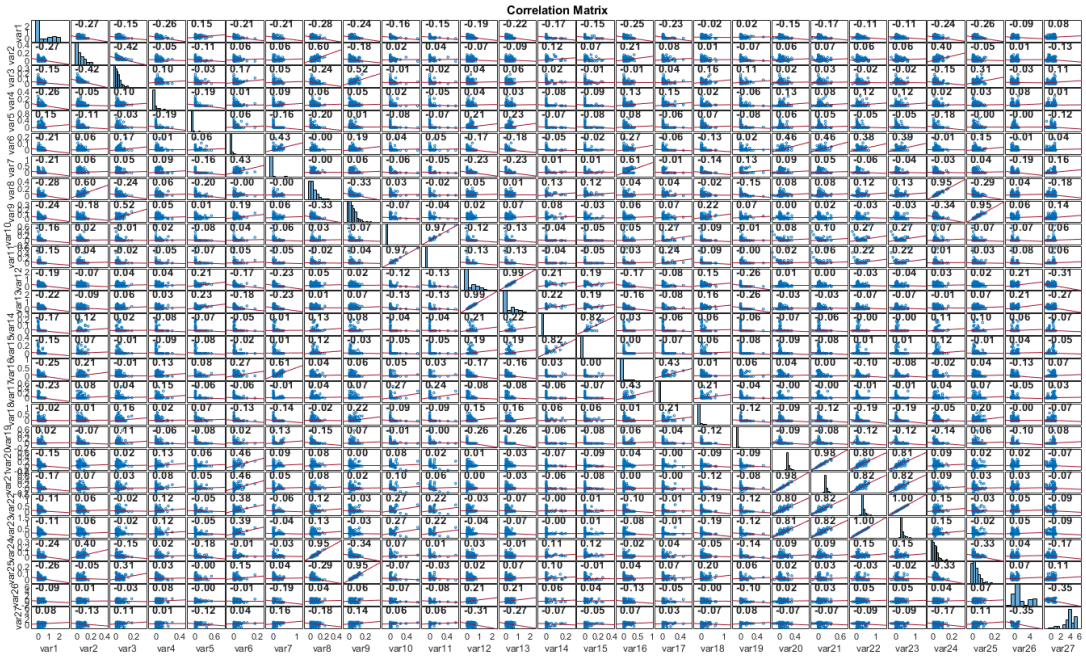


Figure 32. Correlation matrix between the different facial features and the arousal and valence annotations, where var1-var27 represent the facial features as listed in Table 35, excluding Mouth Upper Overlay.

The features' correlation analysis was performed later in the study. It included the original data set (i.e., no processing) collected at the Cyberpsychology laboratory of UQO to understand how the data are correlated and why prediction performances did not improve, regardless of any applied processing steps. The findings of this analysis indicated that the correlations between the features were indeed low.

9.3.3. Importance of Features from Virtual Reality (VR) Data

This section provides a detailed description about various feature ranking algorithms and provides importance scores about the different features.

The minimum redundancy maximum relevance (MRMR) feature ranking algorithm identifies an optimal set of features that are both highly dissimilar from each other and effectively represent the response variable [13,69]. It achieves this by minimizing

redundancy and maximizing relevance, using mutual information to measure both pairwise feature redundancy and feature-response relevance.

The F-Test for feature ranking evaluates the importance of each feature individually through univariate analysis [13]. It tests whether the response values grouped by feature values come from populations with the same mean. A small p-value indicates a significant feature, with the output score being -log(p). Larger scores denote more important features, and a p-value smaller than eps(0) results in an output of Inf. This method works on continuous variables after binning.

The neighbourhood component analysis (NCA) method assigns feature weights using a diagonal adaptation of NCA [13,70]. This non-parametric method aims to maximize prediction accuracy for regression and classification models. It uses regularization to learn feature weights by minimizing an objective function that measures average leave-one-out loss over training data, making it ideal for distance-based supervised models.

The ReliefF feature ranking algorithm calculates feature weights for multiclass categorical variables by penalizing features that give different values to same-class neighbors and rewarding those that differentiate between classes [13,71]. It initializes all feature weights to zero, then iteratively updates them based on the k-nearest samples for each class. RReliefF extends this to continuous variables, using intermediate weights to compute final feature weights [13,72,73]. Both algorithms are effective for distance-based supervised models, with RReliefF using 10 nearest neighbors for ranking features.

The evaluations using MRMR, F-test, NCA, and RReliefF are useful to understand how the data are correlated and why prediction performances did not improve, regardless of any applied processing steps. The scores can help identify which features are more relevant for predicting arousal/valence values, which can be used for feature

selection purposes in future studies. Nonetheless, this is helpful for better data collection as it clarifies what kind of features to focus on.

Table 38 summarizes the feature importance scores for the physiological features with respect to the arousal annotations. The table includes importance scores based on four different ranking algorithms: MRMR, F-Test, NCA, and RReliefF. The ranking order of each feature variable for each ranking algorithm is indicated in the table, where 1 marks the most important feature and 4 marks the least important feature.

Table 38. Importance Scores of Physiological Features with respect to Arousal Annotations

Variable/Feature	MRMR		F-Test		NCA		RReliefF	
	Score	Order	Score	Order	Score	Order	Score	Order
Average HRV: VLF	0.0002	3	0.9764	3	0.6841	4	-0.0030	4
Average HRV: LF	0.0000	4	4.2778	1	0.7426	2	-0.0002	3
Average HRV: HF	0.0077	2	0.7867	4	0.9272	1	-0.0001	2
Heart Rate	0.1796	1	2.1656	2	0.7365	3	0.0066	1

Table 39 summarizes the feature importance scores for the physiological features with respect to the valence annotations. As can be seen, for valence predictions, the physiological features in order of importance are heart rate, average HRV: HF, average HRV: VLF, and average HRV: LF, according to the MRMR and RReliefF tests.

Table 39. Importance Scores of the Physiological Features with respect to Valence Annotations

Variable/Feature	MRMR		F-Test		NCA		RReliefF	
	Score	Order	Score	Order	Score	Order	Score	Order
Average HRV: VLF	0.0082	3	0.5479	4	0.0000	4	0.0007	3
Average HRV: LF	0.0030	4	3.2181	2	1.0336	2	-0.0001	4
Average HRV: HF	0.0608	2	1.2945	3	1.0353	1	0.0027	2
Heart Rate	0.3473	1	5.6432	1	0.7479	3	0.0042	1

Table 40 summarizes the feature importance scores for the pupillometry features with respect to the arousal annotations. As can be seen, for arousal predictions, the pupillometry features in order of importance are pupil X position, pupil Y position, eye openness, and pupil diameter, according to the F-Test and NCA tests. According to the

MRMR and RReliefF tests, the features in order of importance are pupil X position, eye openness, pupil Y position, and pupil diameter, .

Table 40. Importance Scores of the Pupillometry Features with respect to Arousal Annotations

Variable/Feature	MRMR		F-Test		NCA		RReliefF	
	Score	Order	Score	Order	Score	Order	Score	Order
Eye Openness	0.0280	2	0.3270	3	0.000881	3	0.0011	2
Pupil Diameter	0.0000	4	0.1184	4	0.000214	4	-0.0024	4
Pupil X Position	0.0992	1	1.7075	1	0.002419	1	0.0019	1
Pupil Y Position	0.0146	3	0.3600	2	0.000919	2	-0.0011	3

Table 41 summarizes the feature importance scores for the pupillometry features with respect to the valence annotations.

Table 41. Importance Scores of the Pupillometry Features with respect to Valence Annotations

Variable/Feature	MRMR		F-Test		NCA		RReliefF	
	Score	Order	Score	Order	Score	Order	Score	Order
Eye Openness	0.1081	2	2.2454	2	0.0001	3	0.0009	1
Pupil Diameter	0.0000	4	1.7844	3	1.4687	1	-0.0074	4
Pupil X Position	0.1954	1	4.2064	1	0.0001	2	-0.0018	2
Pupil Y Position	0.0195	3	1.0741	4	0.0000	4	-0.0036	3

Table 42 summarizes the feature importance scores for the facial features with respect to the arousal annotations. The ranking order of each feature variable for each ranking algorithm is indicated in the table, where 1 marks the most important feature and 26 marks the least important feature. As shown, the ratings differ from one method to another. These ratings could be used for feature selection purposes to help enhance arousal and valence predictions in the future.

Table 42. Importance Scores of the Facial Features with respect to Arousal Annotations

Variable/Feature	MRMR		F-Test		NCA		RReliefF	
	Score	Order	Score	Order	Score	Order	Score	Order
Jaw Forward	0.1198	3	3.4008	4	0.0001010	7	-0.0121	22
Jaw Right	0.2010	2	0.1338	22	0. 0000304	15	-0.0002	16
Jaw Left	0.0114	20	3.4919	3	0. 0000413	12	0.0041	10
Jaw Open	0.2281	1	0.3888	18	0. 0000165	20	0.0251	1
Mouth Ape Shape	0.0642	6	0.2420	19	0. 0001397	6	-0.0132	23

Variable/Feature	MRMR		F-Test		NCA		RReliefF	
	Score	Order	Score	Order	Score	Order	Score	Order
Mouth O Shape	0.0548	8	0.1316	23	0. 0000134	21	-0.0055	21
Mouth Pout	0.0860	4	4.9393	1	0. 0000094	23	0.0011	14
Mouth Lower Right	0.0158	17	0.0176	25	0. 0000209	18	0.0056	9
Mouth Lower Left	0.0141	18	0.0609	24	0. 0000349	14	0.0075	6
Mouth Smile Right	0.0386	12	0.8174	15	0. 0003351	3	-0.0018	19
Mouth Smile Left	0.0265	14	0.7333	16	0. 0002113	4	-0.0040	20
Mouth Sad Right	0.0550	7	3.5951	2	0. 0005047	2	0.0154	3
Mouth Sad Left	0.0312	13	3.2877	6	0. 0007570	1	0.0159	2
Cheek Puff Right	0.0256	15	3.3230	5	0. 0000854	9	0.0125	5
Cheek Puff Left	0.0184	16	1.5265	10	0. 0000984	8	0.0144	4
Mouth Lower Inside	0.0473	9	2.2434	8	0. 0000071	24	-0.0152	25
Mouth Upper Inside	0.0666	5	1.1952	12	0. 0000529	10	0.0006	15
Mouth Lower Overlay	0.0426	11	1.1347	13	0. 0001483	5	0.0074	7
Mouth Upper Overlay	0	26	0	26	0. 0000440	11	N/A	26
Cheek Suck	0.0444	10	2.0369	9	0. 0000204	19	-0.0150	24
Mouth Lower Right Down	0.0039	23	2.4569	7	0. 0000267	17	-0.0015	18
Mouth Lower Left Down	0.0031	24	0.2022	20	0. 0000021	26	-0.0003	17
Mouth Upper Right Up	0.0129	19	1.1983	11	0. 0000068	25	0.0030	12
Mouth Upper Left Up	0.0069	22	0.1985	21	0. 0000116	22	0.0025	13
Mouth Philtrum Right	0.0010	25	0.4066	17	0. 0000301	16	0.0060	8
Mouth Philtrum Left	0.0106	21	0.8917	14	0. 0000362	13	0.0036	11

Table 43 summarizes the feature importance scores for the facial features with respect to the valence annotations.

Table 43. Importance Scores of the Facial Features with respect to Valence Annotations

Variable/Feature	MRMR		F-Test		NCA		RReliefF	
	Score	Order	Score	Order	Score	Order	Score	Order
Jaw Forward	0.0431	10	2.4022	15	0.0000278	19	-0.0027	19
Jaw Right	0.1898	2	5.5766	5	0. 0000119	24	0.0097	5
Jaw Left	0.0484	9	2.2156	17	0. 0000161	21	-0.0041	22
Jaw Open	0.2072	1	3.4761	9	0. 0000364	18	0.0223	3
Mouth Ape Shape	0.0426	11	1.6718	20	0. 0000738	11	-0.0003	17
Mouth O Shape	0.0323	13	0.0261	25	0. 0000208	20	-0.0022	18
Mouth Pout	0.0237	15	4.2270	7	0. 0001685	4	-0.0031	20
Mouth Lower Right	0.0168	18	3.9624	8	0. 0000057	25	0.0091	6
Mouth Lower Left	0.0030	25	2.4415	14	0. 0000147	22	0.0026	13
Mouth Smile Right	0.0095	24	1.7368	19	0. 0001095	8	0.0033	21
Mouth Smile Left	0.0102	22	2.6963	11	0. 0001297	7	-0.0057	23
Mouth Sad Right	0.0269	14	12.0439	1	0. 0008739	1	0.0289	1
Mouth Sad Left	0.0144	19	7.1450	3	0. 0007652	2	0.0279	2
Cheek Puff Right	0.0795	5	4.7267	6	0. 0000028	26	0.0059	11
Cheek Puff Left	0.0190	16	2.1249	18	0. 0000426	16	0.0087	7
Mouth Lower Inside	0.0099	23	2.5373	12	0. 0001024	9	-0.0169	24
Mouth Upper Inside	0.0118	20	0.6915	22	0. 0000503	15	0.0019	15
Mouth Lower Overlay	0.0333	12	1.3934	21	0. 0003704	3	0.0130	4
Mouth Upper Overlay	0	26	0	26	0. 0000556	12	N/A	26
Cheek Suck	0.0598	6	2.2728	16	0. 0000751	10	-0.0211	25
Mouth Lower Right Down	0.1654	3	10.5653	2	0. 0000552	13	0.0029	12

Variable/Feature	MRMR		F-Test		NCA		RReliefF	
	Score	Order	Score	Order	Score	Order	Score	Order
Mouth Lower Left Down	0.0180	17	5.8446	4	0. 0000530	14	0.0015	16
Mouth Upper Right Up	0.0578	7	0.5812	23	0. 0001408	5	0.0063	9
Mouth Upper Left Up	0.0117	21	0.5174	24	0. 0001339	6	0.0059	10
Mouth Philtrum Right	0.0534	8	2.4485	13	0. 0000135	23	0.0086	8
Mouth Philtrum Left	0.1220	4	2.7869	10	0. 0000369	17	0.0024	14

The findings observed in this section allow to better understand the collected data. Understanding the data is beneficial to improve data collection and/or the regression process in future research.

9.4. Regression over Virtual Reality (VR) Data

As discussed before, sequences of visual data were obtained, enabling the visual data to be treated using three separate methods:

1. Calculating the average of eye and facial features across each sequence to obtain one sample per arousal/valence annotation;
2. Combining the visual sequences into one large data set of samples, where the arousal/valence annotations were duplicated for samples of the same sequence; or
3. Keeping the visual data as sequences, where each sequence was associated with one arousal/valence annotation, and performing sequence-to-label AR.

The first method allowed us to easily synchronize the arousal/valence predictions of physiological and visual data. The second method, on the other hand, required the creation of sequences of predictions which match the original sequences of visual data. Then, average the predictions of each sequence of predictions in order to synchronize them with the predictions on physiological data. The third method also allowed us to synchronize and combine the predictions from the two data modalities, but it led to

lower prediction performances due to the lack of data. Sequence-to-label regression usually demands a lot of data.

In the proof-of-concept studies, optimizable ensemble regression was observed to perform the best with respect to the RMSE, PCC, and CCC performance metrics. Therefore, only an optimizable ensemble regressor was trained and tested on the data collected from VR experiments. An optimizable ensemble regressor was used to predict arousal/valence values from the physiological data, and from visual data in the case of methods 1 and 2 mentioned above. In the case of method 3, a sequence-to-label BiLSTM was used to predict arousal and valence values from sequences of visual data.

The data set was split into training and testing subsets, where 80% were used for training and validation, and 20% were used for testing. As such, the training data set contained 218 samples of physiological data and sequences/samples of visual data, while the testing data set contained 55 samples of physiological data and sequences/samples of visual data. When not averaged, the sequences of visual data corresponded to 78,712 training samples/feature vectors and 18,349 testing samples/feature vectors. 5-fold cross validation was also applied in the case of physiological data and when the visual data were treated as samples. In the case where the visual data were handled as sequences, the training set was divided by another 80-20% split to allow for validation. As a result, the final training set contained 174 sequences, and the validation set contained 44 sequences. The testing set contained 55 sequences.

Table 44 shows the training parameters of the BiLSTM network that was trained and tested on sequences of eye and facial features. Experimentally, the initial learning rate was set to 0.0001, and the number of epochs was set to 30. As there were 174 training sequences, the minimum batch size was set to 3 in order to evenly divide the training

data into 58 equal batches and ensure the whole training set was used during each epoch. This resulted in 58 iterations per epoch (174/3 = 58). The validation frequency was set to 29 to ensure that the training process was validated at least twice per training epoch (58/2 = 29). The number of hidden units was set to 125. Since the network was bidirectional, the number of hidden units multiplied by 2; that is 250 units. Lastly, the SGDM optimizer was used for training.

Table 44. BiLSTM Training Parameters for Sequences of Visual Data

Parameter/Option	Amount/Value
Original Sequences	273
Training Sequences	174
Validation Sequences	44
Testing Sequences	55
Sequence Length	variable
Learning Rate	0.0001
Minimum Batch Size	3
Number of Epochs	30
Iterations per Epoch	174/3 = 58
Validation Frequency	58/2 = 29
Hidden Units	125×2 = 250
Optimizer/Learner	SGDM

9.5. Results on Virtual Reality (VR) Data

As aforementioned, an optimizable ensemble regressor was trained and tested on physiological and visual features collected during VR immersions. The training process was validated using 5-fold cross validation. Additionally, a sequence-to-label BiLSTM was trained and tested on sequences of visual data. This section will further discuss the regression results.

9.5.1. Results on Physiological Virtual Reality (VR) Data

As described in Section 9.4, an optimizable ensemble was tested for the prediction of arousal and valence values from physiological data collected during VR immersions.

Table 45 summarizes the validation and testing prediction performances of this optimizable ensemble regressor in terms of the RMSE, PCC, and CCC performance measures. In the table, the validation results were computed through 5-fold cross-validation over the training data. The testing results were obtained by having the trained model predict arousal and valence values on the testing set of 55 samples. As can be seen, the performance reached at predicting arousal/valence measures from the physiological data samples is low. This is most likely due to the lack of sufficient training data.

Table 45. Prediction Performances on Physiological Data from VR Immersions

Prediction	Regression Type	Ensemble Method	Validation RMSE	Testing RMSE, PCC, CCC
Arousal	Optimizable Ensemble	Bagging	0.58193	0.5386, 0.1283, 0.0229
Valence	Optimizable Ensemble	Bagging	0.44853	0.5012, 0.1233, 0.0102

9.5.2. Results on Visual Virtual Reality (VR) Data

As mentioned previously, the visual data were handled in three separate ways: averaging the samples from each sequence, combining the samples from all sequences, and/or keeping the sequences. After obtaining the averages of sequences and processing them, an optimizable ensemble was tested for the prediction of arousal and valence values from the averages of visual sequences. Table 46 summarizes the validation and testing prediction performances of this optimizable ensemble regressor. As can be seen, the performance reached at predicting arousal/valence measures from the averages of visual sequences is low. This is also likely due to the lack of sufficient training data.

Table 46. Prediction Performances on Averaged Visual Data from VR Immersions

Prediction	Regression Type	Ensemble Method	Validation RMSE	Testing RMSE, PCC, CCC
Arousal	Optimizable Ensemble	Bagging	0.58418	0.5461, -0.0359, -0.0054
Valence	Optimizable Ensemble	Bagging	0.43746	0.4843, 0.2582, 0.1277

Also, an optimizable ensemble was tested for the prediction of arousal and valence values from the combined samples of all visual sequences. Table 47 summarizes the validation and testing prediction performances of this optimizable ensemble regressor. The obtained performance at predicting arousal/valence measures from the combined samples of visual sequences is low. That is due to the use of repeated annotations in this case. Since each sequence was associated with one arousal/valence annotation, when the samples from all sequences were ungrouped and combined, the annotations were repeated.

Table 47. Prediction Performances on Combined Samples of Visual Sequences from VR Immersions

Prediction	Regression Type	Ensemble Method	Validation RMSE	Testing RMSE, PCC, CCC
Arousal	Optimizable	Bagging	0.085182	0.6053, 0.1759, 0.1602
Valence	Ensemble	Bagging	0.076381	0.3761, 0.4330, 0.4112

Finally, a sequence-to-label BiLSTM was trained, validated, and tested on the original sequences of visual data that were collected during the experiments of VR immersions. Table 48 summarizes the validation and testing prediction performances of this BiLSTM network in terms of the RMSE, PCC, and CCC performance measures. In the table, the validation results were computed during the training process using subset of the training sequences. To reiterate previously mentioned information, the training data consisted of 174 sequences of eye and facial features, whereas the validation data consisted of 44 sequences. The testing results were obtained by having the trained neural network predict arousal and valence values on the testing set of 55 sequences. The performance reached at predicting arousal/valence measures from sequences of visual (eye and facial) features is poor. That is most likely due to the lack of sufficient data for training. Neural networks and deep learning, in general, usually require an enormous amount of data for training [68-77]. The required amount of training data varies based on factors such as the type of problem, the complexity of the model, the number of features, and the acceptable error margin. More sophisticated problems

necessitate larger data sets. For instance, deep learning models typically demand millions of data samples, whereas traditional machine learning algorithms might only require thousands of samples. The most common rule is to have 10 times more data samples than features, which is not the case in the collected data set.

Table 48. Prediction Performances on Sequences of Visual Data from VR Immersions

Prediction	Network Type	Validation RMSE	Testing RMSE, PCC, CCC
Arousal	BiLSTM	0.54976	0.5414, 0.0558, 0.0091
Valence		0.46765	0.4998, 0.1339, 0.0408

9.5.3. Results on Multimodal Virtual Reality (VR) Data

In an attempt to improve the prediction performances on the collected VR data, feature and decision fusion were applied to perform multimodal AR. First, the physiological features were fused with the averaged visual sequences of eye and facial features. Then, an optimizable ensemble was trained and tested on the combined feature set. Table 49 displays the prediction performances obtained after feature fusion.

Table 49. Feature Fusion Prediction Performances of the Optimizable Ensemble Jointly Trained on Physiological Samples and Averages of Visual Sequences

Prediction	Regression Type	Ensemble Method	Validation RMSE	Testing RMSE, PCC, CCC
Arousal	Optimizable Ensemble	Bagging	0.58149	0.5445, -0.0205, -0.0022
Valence		Bagging	0.43879	0.4730, 0.3386, 0.1906

Multiple decision fusion schemes were later explored, where the predictions of the optimizable ensemble regressor trained on physiological data were combined with the predictions of the various models that were trained on visual data. Table 50 to Table 52 display the prediction performances after decision fusion of predictions of the optimizable ensemble trained on physiological data with 1) the optimizable ensemble trained on averages of visual sequences, 2) the optimizable ensemble trained on combined samples of visual sequences, and 3) the BiLSTM trained on visual sequences, respectively.

Other machine learning models (i.e., regressors) have also been tested to make sure that the relatively low performance is not due to the inability of the specific solution proposed in this thesis. The prediction performance of the other regressors were poor as well. Please refer to Appendix D for more details about the tested regressors and their performances.

Table 50. Decision Fusion Prediction Performances of the Optimizable Ensembles Trained on Physiological Samples and Averages of Visual Sequences

Prediction	Fused Models	Testing RMSE, PCC, CCC
Arousal	Optimizable Ensemble (Physiological) and	0.5306, 0.2505, 0.1374
Valence	Optimizable Ensemble (Eye/Facial Averages)	0.5060, 0.0440, 0.0198

Table 51. Decision Fusion Prediction Performances of the Optimizable Ensembles Trained on Physiological Samples and Combined Samples from Visual Sequences

Prediction	Fused Models	Testing RMSE, PCC, CCC
Arousal	Optimizable Ensemble (Physiological) and	0.5415, 0.0718, 0.0088
Valence	Optimizable Ensemble (Eye/Facial Samples)	0.4879, 0.2771, 0.0720

Table 52. Decision Fusion Prediction Performances of the Optimizable Ensemble Trained on Physiological Samples and the BiLSTM Trained on Visual Sequences

Prediction	Fused Models	Testing RMSE, PCC, CCC
Arousal	Optimizable Ensemble (Physiological) and	0.5391, 0.1244, 0.0161
Valence	BiLSTM (Eye/Facial Sequences)	0.4988, 0.1607, 0.0259

9.5.4. Oversampling the Virtual Reality (VR) Data Set

In attempt to show that the lack of sufficient data is the source of the poor performance reached, additional data samples were generated from the data set using an improved form of informed oversampling. To do so, the data were first quintuplicated by generating five random subsets from the original data set. Then, white Gaussian noise was added to four of the subsets to generalize the data and make variability between the data samples [83]. One clean (i.e., no noise) subset of data was kept. Lastly, the five subsets were combined to make a larger data set. Note that for this experiment a data set of fused physiological and visual features was used. The corresponding arousal and valence annotations were also quintuplicated using the same steps. Since arousal

and valence values should be between -1 and +1, any noisy value that exceeded these limits was manually modified. Any annotation that went lower that -1 (e.g., -1.583) was changed to -1. Any annotation that went higher than +1 (e.g., +1.583) was changed to +1.

The data set was then processed by applying removal of NaNs, min-max feature scaling, and z-score normalization/standardization (see formulas (5) and (6) in Section 4.1.3). An 80-20% split was then applied to split the data set into two subsets for training and testing, respectively. Thus, a training data set of 1,090 samples (218 × 5) and a testing data set of 275 samples (55 × 5) were created. Each of the training and testing data sets contained 4 physiological and 30 visual features, for a total of 34 features. Finally, an optimizable ensemble regressor was trained and tested to predict arousal and valence values. Table 53 presents the prediction performances on the quintuplicated VR data. As can be seen, quintuplicating the data set significantly improved the prediction performances on arousal and valence predictions. Therefore, one can conclude that the proposed solution(s) is effective for the application of AR in VR, if additional real data were collected.

Table 53. Prediction Performances of the Optimizable Ensemble Trained Quintuplicated Data Set of Fused Physiological and Visual Features

Prediction	Ensemble Method	Validation RMSE	Testing RMSE, PCC, CCC
Arousal	Bagging	0.51815	0.4971, 0.5262, 0.4305
Valence	LSBoost	0.44160	0.4494, 0.4760, 0.4052

To further prove that additional data can improve the performance of the optimizable ensemble regressor, the data set was further decupled using the same steps explained earlier. The decupled data set was also processed through the removal of NaNs, min-max feature scaling, and z-score normalization/standardization. An 80-20% split was applied to the data set to divide it into two subsets for training and testing, respectively. Thus, the produced training data set contained 2,180 samples (218 × 10) and the testing

data set consisted of 550 samples (55 × 10). The training and testing data sets each included a total of 34 features (4 physiological and 30 visual). Finally, an optimizable ensemble regressor was trained and tested to predict arousal and valence values. Table 54 presents the final prediction performances on the decupled VR data. As can be seen, decupling the data set shows an improvement in the prediction performances over those obtained by quintuplicating the data set.

Table 54. Prediction Performances of the Optimizable Ensemble Trained Decupled Data Set of Fused Physiological and Visual Features

Prediction	Ensemble Method	Validation RMSE	Testing RMSE, PCC, CCC
Arousal	Bagging	0.45901	0.4263, 0.7208, 0.5976
Valence	LSBoost	0.40722	0.4318, 0.5882, 0.4738

Figure 33 displays a plot of the predicted arousal and valence values against the actual values using the optimizable ensemble predictions from the quintuplicated data set of physiological and visual features (Table 53) as well as the decupled data set (Table 54).

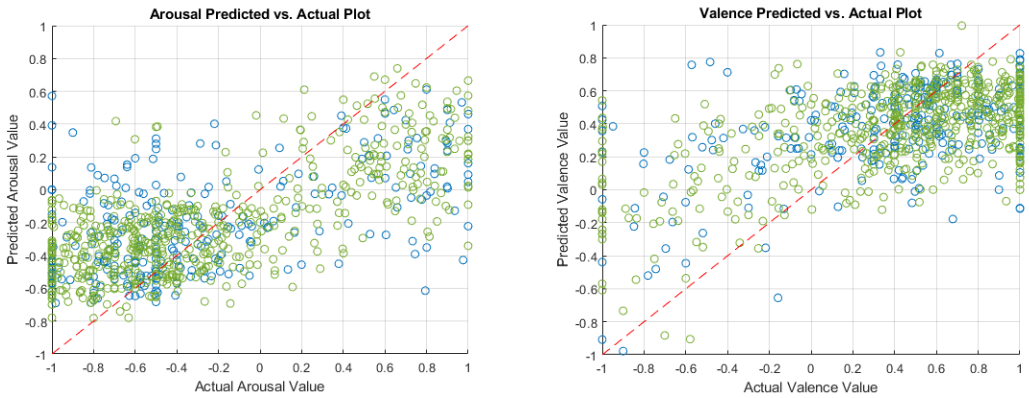


Figure 33. Plots of the predicted versus actual values of (a) arousal and (b) valence by optimizable ensemble regressors using the quintuplicated data set of combined physiological and visual features (blue) and the decupled data set (green).

In this chapter, the performed VR experiments were detailed. The collected data sets were cleaned and processed. Then, the processed data sets were used to train, validate,

and test various types of machine learning regressors to predict arousal and valence values. The initial prediction performances of the tested regressors were not promising. Therefore, more data samples were generated via oversampling. The best prediction performances were achieved by an optimizable ensemble regressor trained on a decupled data set of combined physiological and visual features. For arousal, the reached RMSE was 0.4263, PCC was 0.7208, and CCC was 0.5976. For valence, the obtained RMSE was 0.4318, PCC was 0.5882, and CCC was 0.4738. These prediction performances are coherent with what is in the literature. Nevertheless, data were extracted and preprocessed from a newly collected data set for this study, in which less participants than RECOLA contributed. Even though a direct comparison is not possible, Table 55 compares the prediction performances obtained in this VR study with the prediction performances of state-of-the-art studies on RECOLA. Since this study was conducted on new data with fewer participants, the achieved performances are considered reasonable.

Table 55. Comparison between Prediction Performances Obtained by the VR Study and State-of-the-Art Studies

Prediction	Reference	Technique	RMSE, PCC, CCC
Arousal	**VR Study**	**Optimizable Ensemble**	**0.426, 0.721, 0.598**
	[23]	Decision-level Fusion (BiLSTM)	N/A, N/A, 0.804
	[27]	Random Forests + Linear Regression	0.118, 0.776, 0.762
	[26]	Hierarchical Fusion + Lasso	N/A, N/A, 0.657
	[28]	End2You	N/A, N/A, 0.672
	[29]	Kalman Filters	0.115, 0.774, 0.770
	[34]	Multiple Linear Regressors	N/A, N/A, 0.833
	[36]	SVR + Kalman Filters	N/A, N/A, 0.703
Valence	**VR Study**	**Optimizable Ensemble**	**0.432, 0.588, 0.474**
	[23]	Decision-level Fusion (BiLSTM)	N/A, N/A, 0.528
	[27]	Random Forests + Linear Regression	0.104, 0.634, 0.624
	[26]	Hierarchical Fusion + Multi-task Lasso	N/A, N/A, 0.515
	[28]	End2You	N/A, N/A, 0.521
	[29]	Kalman Filters	0.100, 0.689, 0.687
	[34]	Multiple Linear Regressors	N/A, N/A, 0.596
	[36]	SVR + Kalman Filters	N/A, N/A, 0.681

9.6. Another Perspective on Annotation Labelling

Some experts from the psychology field might argue that the annotation labelling mechanism (see Section 9.1.3) used in this VR study is not adequate. They would suggest modifying the answers to the third and fourth questions from the self-reporting questionnaires. The proposed changes include subtracting 10 from the participants' answers to the third question, only when the answer to the second question is sadness or anger (i.e., a negative emotion). Then, use the existing labelling mechanism (see Section 9.1.3) to estimate valence values. For arousal values, they proposed to treat the answers to the fourth question as standalone, and directly rescale them from -1 to +1 instead of what was done in Section 9.1.3.

The rationale behind the above-mentioned modifications is tied to the way participants might have interpreted the self-reporting questions based on how they are phrased. For instance, the word "energized" in the fourth question might be understood as a positive term, when it can mean energized in a negative manner like being furious. A better term to use in the fourth question would have been "aroused". In terms of the answers to the third question, when participants are asked about how negative they were feeling, some participants might mean that they were feeling the lowest by answering zero, rather than neutral. Hence, experts suggested reversing the numbers in participants' answers. Unlike valence annotations, the arousal annotations do not necessarily depend on the answers to the second question.

Figure 34 shows the correlations matrix for the physiological data combined with the original arousal/valence labels, participants' answers, and modified answers. The modifications did not improve the correlation between the arousal values and physiological data. For valence, the modifications reflected better correlation coefficients, showing a potential impact on prediction performances.

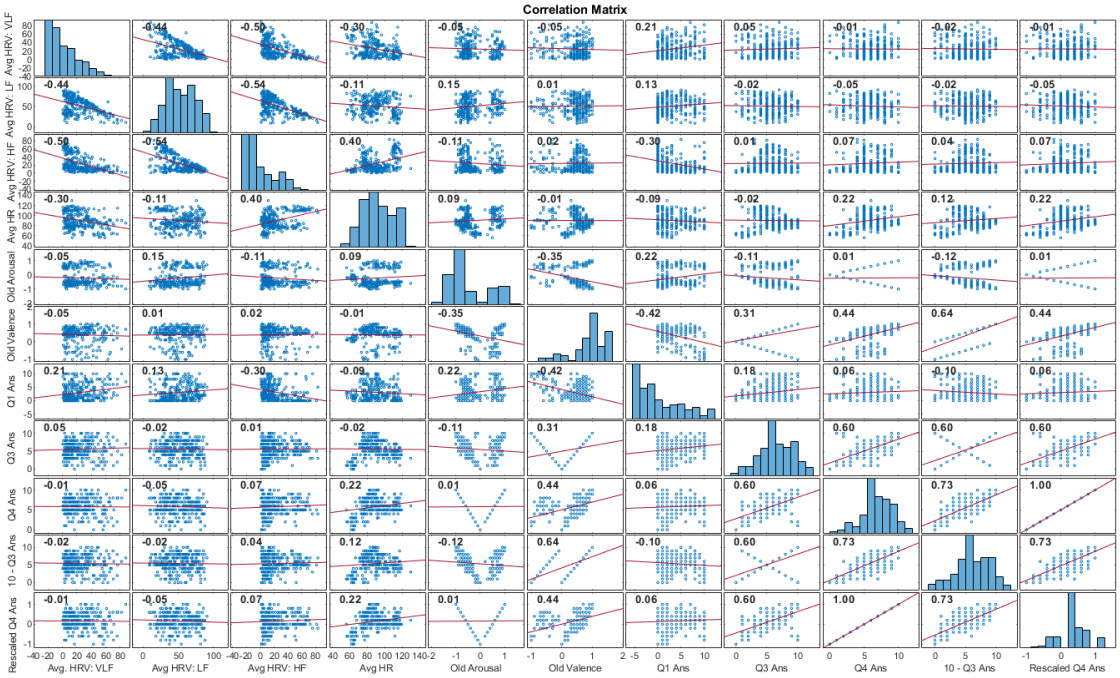


Figure 34. Correlation matrix between the different physiological/ECG features and the modified arousal and valence annotations.

10. CONCLUSIONS AND FUTURE WORK

10.1. Conclusions

This thesis is part of a comprehensive interdisciplinary initiative that includes the Institute of Mental Health Research at the Royal Hospital, the Department of Psychiatry at the University of Ottawa, the Department of Psychoeducation and Psychology, and the Department of Computer Science and Engineering at UQO. The objective is to create and develop a novel, adaptable VR-based intervention to treat cognitive impairments, utilizing integrated methods from computer science and psychology.

I had the pleasure to work in an interdisciplinary team that possesses expertise in machine learning, sensors, VR, cognitive psychology, schizophrenia, and psychiatry. The convergence of these fields offers a unique and unexplored approach to treat cognitive impairments. This research work has the potential to transform the treatments offered, and to have an impact on the lives of patients by reducing cognitive impairment in schizophrenia as well as other brain disorders.

In conclusion, in this thesis, a multi-sensory system was utilized to improve the prediction of mental/cognitive states in terms of the arousal and valence measures. Information from various data sources, collected during VR immersions, were processed to predict arousal and valence values, through machine learning techniques.

The application of AR requires a large amount of data collected from a diverse group of participants. At the initial stage of this thesis, publicly available data sets, specifically RECOLA, were investigated and experimented with as a proof-of-concept mechanism. Continuous emotional predictions (i.e., AR) of the arousal and valence dimensions/measures were performed using features extracted from the EDA and ECG recordings, audio recordings, as well as video recordings of the RECOLA data set. All records of RECOLA were provided as one set. Hence, the data were split into training and testing data sets using an 80-20% split. This approach is common in the engineering field, and in state-of-the-art studies on RECOLA. The data from 18 RECOLA participants were used, totalling in 132,751 samples/images/sequences. 80% of the data were used for training (and validation), whereas 20% were used for testing. Thus, the training data set contained 106,201 samples, and the testing data set consisted of 26,550 samples. The training data set was further split using another 80-20% split for validation purposes in the case of deep learning. As such, 84,960 training samples and 21,241 validation samples were obtained. For classical machine learning, 5-fold cross validation was used. Hyperparameter tuning was conducted in some cases. In the case of classical regressors, such as the optimizable ensemble, no parameter tuning was conducted. Hyperparameter tuning was conducted for training RNNs and CNNs. The tuned parameters included learning rate, minimum batch size, number of epochs, iterations per epoch, validation frequency, hidden units, optimizer/learner type.

The EDA and ECG signals were processed, accompanied with pre-extracted features, and accordingly labelled with their corresponding arousal or valence annotations. Multiple regressors were trained, validated, and tested to predict arousal and valence values. Various preprocessing steps were explored to study their effects on the prediction performance. The replacement of missing values and feature standardization improved the prediction performance. A feature selection mechanism which slightly improved the results on physiological data was also applied. For physiological data,

the best performance was obtained by optimizable ensemble regression. To the best of the author's knowledge, this model outperformed previously published AR studies. This evidence points towards the use of optimizable ensemble regression for physiological data in the following steps of this project.

The video recordings were processed to extract their frames. A face detection algorithm was then applied to the extracted video frames, which were labelled with their corresponding arousal or valence annotations. For images where the algorithm failed to detect a face, manual work was carried out, which can be a time-consuming process and is sometimes prone to error. Pretrained CNNs, such as ResNet-18 and MobileNet-v2, were customized and used to predict the arousal and valence measures. MobileNet-v2 outperformed ResNet-18. Therefore, the MobileNet-v2 CNN will be retained for predicting affective states using visual data in the remainder of this project. The trained MobileNet-v2 CNN was later used to extract deep visual features, which were used to train, validate, and test an optimizable ensemble regressor alone, and after fusion with predesigned visual features. Finally, multimodal decision fusion was applied to fuse the predictions of the optimizable ensemble on the physiological data with the predictions of MobileNet-v2 on the visual data. The achieved multimodal fusion results outperformed results from the literature.

The arousal and valence emotional measures were also predicted using only a portion of the facial features captured in video recordings from the RECOLA data set. Given the nature of HMD in VR systems, the images of detected faces were cropped to only include the lower half of the face. MobileNet-v2 and ResNet-18 were then trained to predict arousal and valence values. Also, an optimizable ensemble regressor was trained using a novel combination of predesigned and deep visual features. The reached results, for both regression methods, were superior to those reported in the literature that used the same data set and complete faces.

Initially, the RECOLA data set was used as a proof-of-concept mechanism. As a continuation of this work, additional data modalities (i.e., acoustic data), additional multimodal fusion techniques were explored, and the impact of their use in the context of a VR system was evaluated. The observed findings were then applied to validate the proposed approach on real data, collected in the Cyberpsychology Laboratory of UQO, as detailed in Chapter 9.

Acoustic features of the RECOLA data set were preprocessed by applying time delay and sequencing, feature standardization/normalization, feature scaling, feature selection, annotation labelling, and data shuffling and splitting. After processing the acoustic features, classical regressors were used for predicting arousal and valence values. The best-performing physiological, visual, and acoustic feature sets were combined to accomplish multimodal AR through feature-level fusion. After feature fusion, classical regressors were used for predicting arousal and valence values.

To validate the findings of the proof-of-concept studies in a VR application, real data during VR immersions of 10 participants have been collected. The collected data were processed using the same approaches as in the proof-of-concept studies. As the size of the initial data set was not sufficient to obtain reasonable prediction performances, an informed oversampling process was used to cope with this problem. Hence, more data samples were created using informed oversampling by quintuplicating and then decupling the data set of physiological and visual features. This has enhanced the prediction performances of the regressors, and hence, proves that the proposed approach is feasible. For arousal predictions, the attained RMSE was 0.4263, PCC was 0.7208, and CCC was 0.5976. For valence predictions, the reached RMSE was 0.4318, PCC was 0.5882, and CCC was 0.4738. These prediction performances are coherent with the literature, given that the VR study was performed on newly collected data with roughly half the number of participants. These performances are anticipated to further

improve with more data. To summarize, Table 56 includes a comparison of the various prediction performances that were obtained in the proof-of-concept studies and VR study of this thesis, as well as the prediction performances of selected state-of-the-art studies.

Table 56. Comparison of Prediction Performances

Data Type	Prediction	Reference	Technique	RMSE, PCC, CCC
Physiological	Arousal	**Chapter 4**	**EDA + ECG Optimizable Ensemble**	**0.0168, 0.997, 0.996**
		[29]	ECG RNN	0.218, 0.407, 0.357
		[29]	EDA RNN	0.250, 0.089, 0.082
	Valence	**Chapter 4**	**EDA + ECG Optimizable Ensemble**	**0.0083, 0.999, 0.998**
		[29]	ECG RNN	0.117, 0.412, 0.364
		[27]	EDA Random Forests	N/A, N/A, 0.206
Visual	Arousal	**Chapter 5**	**Optimizable Ensemble (Full Faces)**	**0.103, 0.850, 0.800**
		Chapter 6	**MobileNet-v2 (Half Faces)**	**0.126, 0.776, 0.772**
		[32]	Ensemble BiLSTM	N/A, N/A, 0.699
	Valence	**Chapter 5**	**Optimizable Ensemble (Full Faces)**	**0.0702, 0.847, 0.805**
		Chapter 6	**MobileNet-v2 (Half Faces)**	**0.0840, 0.765, 0.751**
		[32]	Ensemble BiLSTM	N/A, N/A, 0.617
Acoustic	Arousal	Chapter 7	Optimizable Ensemble	0.154, 0.583, 0.476
		[29]	Linear SVM	0.107, 0.846, 0.846
	Valence	Chapter 7	Optimizable Ensemble	0.114, 0.481, 0.324
		[32]	Ensemble BiLSTM	N/A, N/A, 0.555
Multimodal	Arousal	Chapter 5	Decision Fusion of Optimizable Ensemble (Physiological) + Optimizable Ensemble (Full-Face Visual) Predictions	0.0555, 0.970, 0.948
		Chapter 6	Decision Fusion of Optimizable Ensemble (Physiological) + MobileNet-v2 (Half-Face Visual) Predictions	0.0658, 0.939, 0.934
		Chapter 7	Decision Fusion of MobileNet-v2 (Half-Face Visual) + Optimizable Ensemble (Acoustic) Predictions	0.120, 0.785, 0.725
		Chapter 8	**Feature Fusion and Optimizable Ensemble**	**0.0373, 0.984, 0.979**
		Chapter 9 (VR)	Feature Fusion and Optimizable Ensemble	0.426, 0.721, 0.598
		[34]	Multiple Linear Regressors	N/A, N/A, 0.833
	Valence	Chapter 5	Decision Fusion of Optimizable Ensemble (Physiological) + Optimizable Ensemble (Full-Face Visual) Predictions	0.0373, 0.970, 0.949
		Chapter 6	Decision Fusion of Optimizable Ensemble (Physiological) + MobileNet-v2 (Half-Face Visual) Predictions	0.0441, 0.945, 0.932
		Chapter 7	Decision Fusion of MobileNet-v2 (Half-Face Visual) + Optimizable Ensemble (Acoustic) Predictions	0.0888, 0.752, 0.635
		Chapter 8	**Feature Fusion and Optimizable Ensemble**	**0.0179, 0.993, 0.990**
		Chapter 9 (VR)	Feature Fusion and Optimizable Ensemble	0.432, 0.588, 0.474
		[29]	Kalman Filters	0.100, 0.689, 0.687

10.2. Challenges and Limitations

As observed, one factor that directly affected the results in the final VR study was the lack of data and data discrepancies in terms of sampling rates. Another factor that could have affected the results of affect recognition in the final VR study was the reliability of the arousal and valence annotations. The arousal and valence annotations in RECOLA were based on both self-reporting and the expert opinion of six professional raters. In the collected data for the VR study, arousal and valence annotations were solely based on the participants' answers to self-reporting questions.

The machine learners in this thesis were all trained and tested on a computer equipped with a NVIDIA GeForce GTX 1070 graphics processing unit (GPU). Using a newer GPU is recommended to optimize training speeds. Please refer to Appendix E for more information on computational load and complexity of the implemented machine learning models.

10.3. Contributions

As a proof-of-concept, the RECOLA data set was exploited to predict the arousal and valence values of affective states using physiological and visual data. Various processing methods were applied to enhance the quality of data, and improve the prediction performance of multiple regressors and CNNs. The following were achieved in this thesis: 1) a novel combination of processing steps was proposed to enable the synchronous use of physiological, visual, and acoustic data modalities; 2) deep facial features were extracted for AR by customizing the MobileNet-v2 CNN; 3) unique combinations of physiological, visual (predesigned and deep), and/or acoustic features were fused; 4) an outliers-based feature selection mechanism was proposed to help optimize the prediction of the regressors; 5) a unimodal and multimodal decision fusion

mechanism was introduced to combine the AR predictions; 6) feature fusion and multimodal AR were studied; 7) a research protocol was submitted and approval from REB or IRB was obtained; 8) physiological and visual data were extracted and preprocessed from data collected during VR immersions; and 9) AR was performed on data collected during VR immersions.

The research work performed in this thesis has been published in five journals [78-82]:

- [78] Joudeh, I.O.; Cretu, A.-M.; Guimond, S.; Bouchard, S. Prediction of Emotional Measures via Electrodermal Activity (EDA) and Electrocardiogram (ECG). *Engineering Proceedings* **2022**, *27*, 47.

- [79] Joudeh, I.O.; Cretu, A.-M.; Bouchard, S.; Guimond, S. Prediction of Continuous Emotional Measures through Physiological and Visual Data. *Sensors* **2023**, *23*, 5613.

- [80] Joudeh, I.O.; Cretu, A.-M.; Bouchard, S.; Guimond, S. Prediction of Emotional States from Partial Facial Features for Virtual Reality Applications. *Annual Review of Cybertherapy and Telemedicine* **2023**, *21*.

- [81] Joudeh, I.O.; Cretu, A.-M.; Bouchard, S. Optimizable Ensemble Regression for Arousal and Valence Predictions from Visual Features. *Engineering Proceedings* **2023**, *58*, 3.

- [82] Joudeh, I.O.; Cretu, A.-M.; Bouchard, S. Predicting the Arousal and Valence Values of Emotional States Using Learned, Predesigned, and Deep Visual Features. *Sensors* **2024**, *24*, 4398.

10.4. Future Work

Future research, extending beyond this thesis, could investigate the integration of additional sensors to not only predict affective states, but also to assess cognitive effort during VR interventions. This would facilitate the creation of more tailored and effective treatments for individuals with cognitive impairments. Additionally, practical methods could be developed to use the detected emotional states to adjust feedback, aiding users in specific tasks. For instance, this could involve controlling the amount of information displayed through selectively densified objects, providing extra task support by limiting the number of possible movements in 3D to only those necessary for the task, or restricting movements to one type or direction at a time, and offering assistance in unexpected situations.

Other ways to process the data, other machine learning algorithms, other feature fusion or feature selection methods, or other decision fusion methods could be examined. Data collection through additional sensors and sources of data (e.g., EDA/GSR, audio, etc.), during VR immersions, could be explored as well.

Furthermore, the data set that was collected during the VR experiments includes video recordings of the facial expressions of participants. Those recordings are a potential source for future contributions. One could also investigate the possibility to improve the data collection process if any additional VR tests will be performed. Perhaps, propose a better mechanism to collect arousal and valence annotations in a continuous and more reliable manner. Other authors should also consider spending more time to collect continuous physiological signals and features.

APPENDIX A: PRELIMINARY SUBJECT-BASED RESULTS ON PHYSIOLOGICAL AND VISUAL DATA FROM RECOLA

This appendix includes tables summarizing all prediction performances that were obtained using the physiological and visual recordings of RECOLA during subject-based regression. Various regressors were trained on physiological data (EDA and ECG) for 15 RECOLA subjects, labelled with arousal or valence values. The training process was validated using 5-fold cross validation. Then, the trained regressors were tested on physiological data from 3 different RECOLA subjects. Table 57 shows the breakdown of the physiological data that were used for training, validation, and testing. Table 58 and Table 59 present the arousal and valence prediction performances for subject-based regression using physiological data.

Table 57. Breakdown of RECOLA's Physiological Data for Subject-based Regression

Parameters and Options	EDA Samples	ECG Samples
Training	112,072	112,072
Validation	5-fold cross validation	
Testing	22,447	22,447

Table 58. Arousal Prediction Performances of Subject-based Regression using RECOLA's Physiological Data

Data Set	Data Nature	Regression Type	Prediction	Validation RMSE	Testing RMSE, PCC, CCC
RECOLA	EDA, ECG	Linear Regression	arousal	0.19473	0.1914, -0.0962, -0.0040
		Interactions Linear Regression		0.19472	0.1914, -0.0980, -0.0041
		Robust Linear Regression		0.19482	0.1910, -0.0907, -0.0030
		Stepwise Linear Regression		0.19472	0.1914, -0.0980, -0.0041
		Fine Tree		0.18681	0.2533, -0.0812, -0.0790
		Medium Tree		0.1713	0.2402, -0.0906, -0.0842
		Coarse Tree		0.16551	0.2331, -0.0943, -0.0840
		Optimizable Tree		0.16539	0.2312, -0.1063, -0.0927
		Linear SVM		0.19612	0.1910, -8.0217e-04, -2.7534e-05
		Quadratic SVM		0.20726	0.1925, 0.0161, 0.0013
		Cubic SVM		1.5025	0.9435, 0.0682, 0.0016
		Fine Gaussian SVM		0.18475	0.2211, -0.1656, -0.1207
		Medium Gaussian SVM		0.19078	0.1957, -0.0064, -0.0025
		Coarse Gaussian SVM		0.19415	0.1917, -4.9421e-04, -1.0202e-04
		Optimizable SVM		0.1951	0.1907, -3.0601e-16, -3.5329e-30
		Ensemble: Boosted Trees		0.17514	0.1958, -0.1194, -0.0298
		Ensemble: Bagged Trees		0.16429	0.2225, -0.0942, -0.0768
		Optimizable Ensemble		0.16123	0.2312, -0.1076, -0.0937
		Narrow Neural Network		0.18693	0.1969, -0.0917, -0.0298
		Medium Neural Network		0.18248	0.2212, -0.1697, -0.1251
		Wide Neural Network		0.17914	0.2138, -0.1717, -0.1100
		Bi-layered Neural Network		0.18155	0.2124, -0.1672, -0.1046
		Tri-layered Neural Network		0.18053	0.2122, -0.1571, -0.0990
		Optimizable Neural Network		0.18359	0.1981, -0.0566, -0.0208
		SVM Kernel		0.18455	0.2209, -0.1570, -0.1132
		Least Squares Regression Kernel		0.18147	0.2080, -0.1819, -0.0969

Table 59. Valence Prediction Performances of Subject-based Regression using RECOLA's Physiological Data

Data Set	Data Nature	Regression Type	Prediction	Validation RMSE	Testing RMSE, PCC, CCC
RECOLA	EDA, ECG	Linear Regression	valence	0.13862	0.0953, -0.0154, -0.0032
		Interactions Linear Regression		0.13862	0.0953, -0.0152, -0.0032
		Robust Linear Regression		0.13919	0.0947, -0.0162, -0.0033
		Stepwise Linear Regression		0.13862	0.0953, -0.0154, -0.0032
		Fine Tree		0.1325	0.1354, -0.0260, -0.0251
		Medium Tree		0.12136	0.1236, -0.0359, -0.0332
		Coarse Tree		0.1173	0.1186, -0.0462, -0.0404
		Optimizable Tree		0.11704	0.1176, -0.0520, -0.0446
		Linear SVM		0.14051	0.0951, -0.0167, -0.0030
		Quadratic SVM		0.15699	0.0977, -0.0379, -0.0123
		Cubic SVM		0.99517	2.8471, 0.0708, 1.1316e-04
		Fine Gaussian SVM		0.12592	0.1022, -0.0882, -0.0537
		Medium Gaussian SVM		0.13245	0.0978, 0.0188, 0.0097
		Coarse Gaussian SVM		0.13827	0.0961, -0.0380, -0.0123
		Optimizable SVM		0.14208	0.0971, 4.5891e-16, 3.9616e-29
		Ensemble: Boosted Trees		0.1227	0.0958, 0.0716, 0.0327
		Ensemble: Bagged Trees		0.11609	0.1088, -6.4659e-04, -5.1916e-04
		Optimizable Ensemble		0.11566	0.1086, -0.0343, -0.0250
		Narrow Neural Network		0.13204	0.0979, 0.0572, 0.0330
		Medium Neural Network		0.12603	0.0970, 0.0668, 0.0354
		Wide Neural Network		0.1234	0.0980, 0.0834, 0.0472
		Bi-layered Neural Network		0.12463	0.0993, 0.0430, 0.0264
		Tri-layered Neural Network		0.12384	0.0988, 0.0376, 0.0215
		Optimizable Neural Network		0.12676	0.1011, -0.0265, -0.0148
		SVM Kernel		0.12821	0.0952, 0.0868, 0.0431
		Least Squares Regression Kernel		0.12591	0.1057, -0.1701, -0.0867

Various regressors were trained on physiological data and statistical features for 15 RECOLA subjects, labelled with arousal or valence values. The training process was validated using 5-fold cross validation. Then, the trained regressors were tested on the physiological data and statistical features from 3 different RECOLA subjects. Table 60 shows the breakdown of the physiological data with statistical features that were used for training, validation, and testing.

Table 60. Breakdown of RECOLA's Physiological Data alongside Statistical Features for Subject-based Regression

Parameters and Options	EDA Samples	ECG Samples	EDA Features	ECG Features	Final Dimensions
Training	112,072	112,072	24	8	112,072 × 32
Validation	5-fold cross validation				
Testing	22,447	22,447	24	8	22,447 × 32

Table 61 and Table 62 present the arousal and valence prediction performances for subject-based regression using physiological data and statistical features.

Table 61. Arousal Prediction Performances of Subject-based Regression using RECOLA's Physiological Data alongside Statistical Features

Data Set	Data Nature	Regression Type	Prediction	Validation RMSE	Testing RMSE, PCC, CCC
RECOLA	EDA and ECG features	Fine Tree	arousal	0.18766	0.2640, -0.0689, -0.0685
		Medium Tree		0.17152	0.2480, -0.0911, -0.0874
		Coarse Tree		0.16253	0.2379, -0.1011, -0.0925
		Optimizable Tree		0.16163	0.2338, -0.1111, -0.0987
		Ensemble: Boosted Trees		0.17307	0.1983, -0.0953, -0.0302
		Ensemble: Bagged Trees		0.14802	0.2311, -0.1277, -0.1099
		Optimizable Ensemble		0.14460	0.2325, -0.1337, -0.1158

Table 62. Valence Prediction Performances of Subject-based Regression using RECOLA's Physiological Data alongside Statistical Features

Data Set	Data Nature	Regression Type	Prediction	Validation RMSE	Testing RMSE, PCC, CCC
RECOLA	EDA and ECG features	Fine Tree	valence	0.13439	0.1482, -0.0475, -0.0463
		Medium Tree		0.12151	0.1318, -0.0409, -0.0399
		Coarse Tree		0.11565	0.1210, -0.0377, -0.0352
		Optimizable Tree		0.115	0.1183, -0.0394, -0.0359
		Optimizable SVM		0.14209	0.0971, -9.8002e-16, -6.7790e-30
		Ensemble: Boosted Trees		0.12169	0.0961, 0.0631, 0.0314
		Ensemble: Bagged Trees		0.10596	0.1160, -0.0398, -0.0347
		Optimizable Ensemble		0.10545	0.1146, -0.0486, -0.0413

Various regressors were trained on RECOLA's physiological data and predesigned features of 15 subjects, labelled with arousal or valence values. The training process was validated using 5-fold cross validation. Then, the trained regressors were tested on the physiological data and features from 3 different RECOLA subjects. Table 63 shows the breakdown of RECOLA's physiological data and features that were used for training, validation, and testing.

Table 63. Breakdown of RECOLA's Physiological Data and Predesigned Features for Subject-based Regression

Parameters and Options	EDA Samples	ECG Samples	EDA Features	ECG Features	Final Dimensions
Training	110,599	110,599	63	55	110,599 × 118
Validation	5-fold cross validation				
Testing	22,152	22,152	63	55	22,152 × 118

Table 64 displays the arousal prediction performances for subject-based regression using RECOLA's physiological data and predesigned features.

Table 64. Arousal Prediction Performances of Subject-based Regression using RECOLA's Physiological Data and Predesigned Features

Data Set	Data Nature	Validation	Regression Type	Prediction	Validation RMSE	Testing RMSE, PCC, CCC
RECOLA	EDA, ECG	5-fold cross validation	Fine Tree	arousal	0.038851	0.2568, -0.0063, -0.0061
			Medium Tree		0.04583	0.2425, 0.0072, 0.0069
			Coarse Tree		0.065707	0.2423, 0.0118, 0.0115
			Optimizable Tree		0.036528	0.2465, -0.0141, -0.0137
			Optimizable SVM		0.19603	0.1916, 1.9147e-16, 1.6591e-30
			Ensemble: Boosted Trees		0.15629	0.2176, -0.1914, -0.1230
			Ensemble: Bagged Trees		0.021601	0.2177, -0.1529, -0.0920

Various regressors were trained on RECOLA's physiological sequences of 15 subjects, labelled with arousal or valence values. The training process was validated using an 80-20% split validation. Then, the trained regressors were tested on the physiological sequences of 3 different RECOLA subjects. Table 65 shows the breakdown of the sequences that were used for training, validation, and testing.

Table 65. Breakdown of RECOLA's Physiological Sequences used for Subject-based Regression

Parameters and Options	EDA Samples	ECG Samples
Training	89,658	89,658
Validation (80-20% Split)	22,414	22,414
Testing	22,447	22,447
Learning Rate	0.0001	
Minimum Batch Size	9	
Number of Epochs	30	
Iterations per Epoch	89658/9 = 9,962	
Validation Frequency	9962/2 = 4981	

Table 66 displays the arousal and valence prediction performances for subject-based regression using physiological sequences from RECOLA.

Table 66. Arousal and Valence Prediction Performances of Subject-based Regression using Physiological Sequences of RECOLA

Data Set	Data Nature	Network Type	Hidden Units	Epochs	Learning Rate	Learner/ Optimizer	Prediction	Validation RMSE	Testing RMSE, PCC, CCC
RECOLA	EDA, ECG **(Not Normalized)**	LSTM	200	30	0.0001	SGDM	Arousal, valence	0.22493	**Arousal:** 0.1953, -0.0887, -0.0245 **Valence:** 0.1003, 0.3294, 0.1500
	EDA, ECG **(Not Normalized)**	LSTM	125	30	0.0001	SGDM	Arousal, valence	0.22624	**Arousal:** 0.1952, -0.0779, -0.0214 **Valence:** 0.1039, 0.3327, 0.2012
	EDA, ECG **(Not Normalized)**	LSTM	100	30	0.0001	SGDM	Arousal, valence	0.22562	**Arousal:** 0.1964, -0.1019, -0.0272 **Valence:** 0.1017, 0.3260, 0.2045
	EDA, ECG **(Not Normalized)**	LSTM	200	30	0.001	SGDM	Arousal, valence	0.22648	**Arousal:** 0.1954, -0.1222, -0.0291 **Valence:** 0.0993, 0.3268, 0.1543
	EDA, ECG **(Normalized)**	LSTM	200	30	0.0001	SGDM	Arousal, valence	0.68078	**Arousal:** 0.5647, -0.1307, -0.0017 **Valence:** 0.3856, 0.3533, 0.0089
	EDA, ECG **(Not Normalized)**	BiLSTM	100, 100	30	0.0001	SGDM	Arousal	0.18739	0.1980, -0.1207, -0.0304
	EDA, ECG	BiLSTM	100, 100	150	0.0001	SGDM	Arousal	0.18615	0.1938, -0.0364, -0.0104

Data Set	Data Nature	Network Type	Hidden Units	Epochs	Learning Rate	Learner/ Optimizer	Prediction	Validation RMSE	Testing RMSE, PCC, CCC
	(Not Normalized)								
	EDA, ECG **(Not Normalized)**	BiLSTM	100, 100	150	0.0001	SGDM	Arousal, valence	0.22304	**Arousal:** 0.1958, -0.0995, -0.0289 **Valence:** 0.1038, 0.3260, 0.1712

CNN(s) were trained on video frames from 20 RECOLA videos/subjects, labelled with arousal values. The trained CNN(s) were then tested on frames from 3 RECOLA different videos/subjects. Table 67 shows the breakdown of the frames were used for training, validation, and testing.

Table 67. Breakdown of RECOLA's Video Frames for Subject-based CNN Regression

Parameters and Options	Original	80-20% Split	70-30% Split
Training Frames	98,336	78,669	68,836
Validation Frames	N/A	19,667	29,500
Testing Frames	15,288	15,288	15,288
Learning Rate		various	various
Minibatch Size		9	4
Number of Epochs		10	10
Iterations per Epoch		78669/9 = 8,741	68836/4 = 17,209
Validation Frequency		8741/3 = 2,913	17209/3 = 5,736

Table 68 and Table 69 display the arousal and valence prediction performances for subject-based regression using various schemes of RECOLA's video frames.

Table 68. Arousal and Valence Prediction Performances of Subject-based MobileNet Regression using the Video Frames of RECOLA

Data Set	Data Nature	Validation	CNN Type	Learner/ Optimizer	Learning Rate	Prediction	Validation RMSE	Testing RMSE, PCC, CCC
RECOLA	full faces	80-20% split	MobileNet	SGDM	0.0001	Arousal	0.18669	0.17177
	half faces	80-20% split	MobileNet	SGDM	0.0001	Arousal	0.21955	0.21235
	full faces	80-20% split	MobileNet	SGDM	0.001+	Arousal	NaN	NaN
	half faces	80-20% split	MobileNet	SGDM	0.001+	Arousal	NaN	NaN
	half face	80-20% split	MobileNet	SGDM	0.0001	Arousal	0.19969	0.1636, 0.1115, 0.1029
	half face	80-20% split	MobileNet	SGDM	0.0001	Valence	0.12756	0.1133, 0.0689, 0.0654
	half face	80-20% split	MobileNet	SGDM	0.0001	Arousal, Valence	0.22933	**Arousal:** 0.1748, 0.0164, 0.2043 **Valence:** 0.1118, 0.3858, 0.2130
	half face	80-20% split	MobileNet	ADAM	0.0001	Arousal	0.20321	0.1768, 0.1061, 0.1006
	half face	80-20% split	MobileNet	ADAM	0.0001	Valence	0.14332	0.1330, 0.1532, 0.1433
	half face	80-20% split	MobileNet	ADAM	0.0001	Arousal, Valence	0.2671	**Arousal:** 0.1786, 0.0343, 0.0228 **Valence:** 0.1693, 0.0434, 0.0130

Table 69. Arousal and Valence Prediction Performances of Subject-based ResNet and MobileNet Regression using the Video Frames of RECOLA

Data Set	Data Nature	Validation	CNN Type	Learner/ Optimizer	Learning Rate	Prediction	Validation RMSE	Testing RMSE, PCC, CCC
RECOLA	half face	80-20% split	ResNet18	SGDM	0.0001	Arousal	0.16927	0.1632, 0.1482, 0.1408
	half face	80-20% split	ResNet18	SGDM	0.0001	Valence	0.11570	0.1133, 0.2981, 0.2500
	half face	80-20% split	ResNet18	SGDM	0.0001	Arousal, Valence	0.24147	**Arousal:** 0.2044, 0.1297, 0.0746 **Valence:** 0.0990, 0.5244, 0.4740
	half face	80-20% split	ResNet50	SGDM	0.0001	Arousal	0.16541	0.1762, 0.1323, 0.1087
	half face	80-20% split	ResNet50	SGDM	0.0001	Valence	0.13157	0.1152, 0.3045, 0.2717
	half face	80-20% split	ResNet50	SGDM	0.0001	Arousal, Valence	0.18361	**Arousal:** 0.1478, 0.2227, 0.1984 **Valence:** 0.0939, 0.4855, 0.4572
	half face	80-20% split	ResNet101	SGDM	0.0001	Arousal	0.16166	0.1640, 0.2298, 0.1918
	half face	80-20% split	ResNet101	SGDM	0.0001	Valence	0.10592	0.0977, 0.3473, 0.3239
	half face	80-20% split	ResNet101	SGDM	0.0001	Arousal, Valence	0.18389	**Arousal:** 0.1514, 0.2276, 0.1995 **Valence:** 0.0952, 0.4874, 0.4453
	half face	N/A	Average ResNet (ResNet18, ResNet50, ResNet101)	N/A	N/A	Arousal, Valence	N/A	**Arousal:** 0.1447, 0.2489, 0.1867 **Valence:** 0.0865, 0.5608, 0.4867
	half face	80-20% split	MobileNet	SGDM	0.0001	Arousal	0.16711	0.1627, 0.0941, 0.0846
	half face	80-20% split	MobileNet	SGDM	0.0001	Valence	0.11223	0.1092, 0.1693, 0.1510
	half face	80-20% split	MobileNet	SGDM	0.0001	Arousal, Valence	0.19866	**Arousal:** 0.1586, 0.1492, 0.1274 **Valence:** 0.1020, 0.4927, 0.4477

Data Set	Data Nature	Validation	CNN Type	Learner/ Optimizer	Learning Rate	Prediction	Validation RMSE	Testing RMSE, PCC, CCC
	half face	N/A	MobileNet + Average ResNet	N/A	N/A	Arousal, Valence	N/A	**Arousal:** 0.1459, 0.2301, 0.1703 **Valence:** 0.0904, 0.5523, 0.4643
	half face	N/A	ResNet18, ResNet50, ResNet101, MobileNet	N/A	N/A	Arousal, Valence	N/A	**Arousal:** 0.1442, 0.2447, 0.1768 **Valence:** 0.0876, 0.5602, 0.4702
	half face	80-20% split	NasNet	SGDM	0.0001	Arousal	0.2152	0.1917, 0.0371, 0.0370

APPENDIX B: DETAILED RESULTS ON PHYSIOLOGICAL DATA

This appendix includes tables summarizing all prediction performances that were obtained using the physiological recordings (EDA and ECG signals) of RECOLA. Table 70 presents the arousal prediction performances obtained by applying different combinations of the processing steps discussed earlier in this thesis, such as the removal of missing values (i.e., NaNs), feature scaling, and feature standardization.

Table 70. Arousal Prediction Performances using RECOLA's Physiological Data with Differing Processing Scenarios

Scaled?	Standardized (z-score)?	Has NaNs?	Prediction	Regressor	Validation RMSE	Testing RMSE, PCC, CCC
No	No	Yes	arousal	Fine Tree	0.048101	0.0429, 0.9740, 0.9739
				Medium Tree	0.056231	0.0502, 0.9642, 0.9640
				Coarse Tree	0.077392	0.0698, 0.9292, 0.9276
				Optimizable Tree	0.046698	0.0405, 0.9769, 0.9769
				Boosted Trees	0.14376	0.1440, 0.6811, 0.5234
				Bagged Trees	0.028162	0.0233, 0.9941, 0.9918
				Optimizable Ensemble	**0.021198**	**0.0173, 0.9966, 0.9956**
No	No	No	arousal	Fine Tree	0.047736	0.0420, 0.9751, 0.9751
				Medium Tree	0.055758	0.0494, 0.9653, 0.9651
				Coarse Tree	0.076991	0.0697, 0.9296, 0.9279
				Optimizable Tree	0.045999	0.0396, 0.9780, 0.9779
				Boosted Trees	0.14362	0.1440, 0.6821, 0.5237

Scaled?	Standardized (z-score)?	Has NaNs?	Prediction	Regressor	Validation RMSE	Testing RMSE, PCC, CCC
				Bagged Trees	0.027193	0.0225, 0.9946, 0.9925
				Optimizable Ensemble	0.055619	0.0532, 0.9596, 0.9584
Yes	No	No	arousal	Fine Tree	0.047711	0.0420, 0.9751, 0.9750
				Medium Tree	0.055753	0.0495, 0.9652, 0.9651
				Coarse Tree	0.076989	0.0697, 0.9296, 0.9279
				Optimizable Tree	0.045976	0.0396, 0.9779, 0.9779
				Boosted Trees	0.14362	0.1440, 0.6821, 0.5237
				Bagged Trees	**0.027009**	**0.0225, 0.9946, 0.9925**
				Optimizable Ensemble	0.027166	0.0237, 0.9926, 0.9918
No	Yes	No	arousal	Fine Tree	0.046975	0.0420, 0.9751, 0.9751
				Medium Tree	0.055288	0.0494, 0.9653, 0.9651
				Coarse Tree	0.076615	0.0697, 0.9296, 0.9279
				Optimizable Tree	0.045176	0.0396, 0.9780, 0.9779
				Boosted Trees	0.14359	0.1440, 0.6821, 0.5237
				Bagged Trees	0.02689	0.0221, 0.9948, 0.9927
				Optimizable Ensemble	**0.021043**	**0.0170, 0.9967, 0.9958**
Yes	Yes	No	arousal	Fine Tree	0.04771	0.0420, 0.9751, 0.9750
				Medium Tree	0.05575	0.0494, 0.9653, 0.9651
				Coarse Tree	0.076989	0.0697, 0.9296, 0.9279
				Optimizable Tree	0.045975	0.0396, 0.9779, 0.9779
				Boosted Trees	0.14362	0.1440, 0.6821, 0.5237
				Bagged Trees	0.026945	0.0227, 0.9945, 0.9923
				Optimizable Ensemble	**0.023379**	**0.0194, 0.9954, 0.9945**

Table 71 presents the valence prediction performances obtained by applying different combinations of the processing steps discussed earlier in this thesis, such as the removal of missing values (i.e., NaNs), feature scaling, and feature standardization.

Table 71. Valence Prediction Performances using RECOLA's Physiological Data with Differing Processing Steps

Scaled?	Standardized (z-score)?	Has NaNs?	Prediction	Regressor	Validation RMSE	Testing RMSE, PCC, CCC
No	No	Yes	valence	Fine Tree	0.028757	0.0248, 0.9814, 0.9813
				Medium Tree	0.033463	0.0288, 0.9748, 0.9747
				Coarse Tree	0.047371	0.0429, 0.9431, 0.9422
				Optimizable Tree	0.027821	0.0243, 0.9821, 0.9821
				Boosted Trees	0.10128	0.1003, 0.6669, 0.4924
				Bagged Trees	0.018977	0.0156, 0.9943, 0.9921
				Optimizable Ensemble	**0.014697**	**0.0126, 0.9957, 0.9950**
No	No	No	valence	Fine Tree	0.026639	0.0237, 0.9829, 0.9829
				Medium Tree	0.032055	0.0281, 0.9760, 0.9760
				Coarse Tree	0.046648	0.0422, 0.9449, 0.9442
				Optimizable Tree	0.025588	0.0230, 0.9840, 0.9840
				Boosted Trees	0.099214	0.0987, 0.6784, 0.5164
				Bagged Trees	0.016937	0.0136, 0.9959, 0.9940
				Optimizable Ensemble	**0.012208**	**0.0105, 0.9970, 0.9965**
Yes	No	No	valence	Fine Tree	0.027735	0.0237, 0.9829, 0.9829
				Medium Tree	0.032494	0.0281, 0.9760, 0.9760
				Coarse Tree	0.046696	0.0422, 0.9449, 0.9442
				Optimizable Tree	0.026567	0.0230, 0.9840, 0.9840
				Boosted Trees	0.099295	0.0987, 0.6784, 0.5164
				Bagged Trees	0.016949	0.0137, 0.9958, 0.9940
				Optimizable Ensemble	**0.014091**	**0.0115, 0.9965, 0.9958**
No	Yes	No	valence	Fine Tree	0.026728	0.0237, 0.9830, 0.9830
				Medium Tree	0.031906	0.0280, 0.9761, 0.9760
				Coarse Tree	0.046718	0.0422, 0.9450, 0.9443
				Optimizable Tree	0.025805	0.0230, 0.9840, 0.9840
				Boosted Trees	0.099252	0.0987, 0.6784, 0.5164
				Bagged Trees	0.016891	0.0137, 0.9958, 0.9940
				Optimizable Ensemble	**0.012907**	**0.0106, 0.9971, 0.9965**
Yes	Yes	No	valence	Fine Tree	0.027736	0.0237, 0.9830, 0.9830

Scaled?	Standardized (z-score)?	Has NaNs?	Prediction	Regressor	Validation RMSE	Testing RMSE, PCC, CCC
				Medium Tree	0.032494	0.0281, 0.9760, 0.9760
				Coarse Tree	0.046696	0.0422, 0.9449, 0.9442
				Optimizable Tree	0.026569	0.0229, 0.9841, 0.9841
				Boosted Trees	0.099296	0.0987, 0.6784, 0.5164
				Bagged Trees	0.016800	0.0135, 0.9959, 0.9942
				Optimizable Ensemble	**0.012570**	**0.0104, 0.9971, 0.9966**

Table 72 presents the arousal and valence prediction performances obtained by applying feature selection, with 5% of outliers.

Table 72. Prediction Performances using RECOLA's Physiological Data, after Feature Scaling, Standardization, and 5% Feature Selection

Scaled?	Standardized (z-score)?	Has NaNs?	Feature Selection	Prediction	Regressor	Validation RMSE	Testing RMSE, PCC, CCC
Yes	Yes	No	5% Outliers (50 features)	arousal	Fine Tree	0.044788	0.0393, 0.9783, 0.9782
					Medium Tree	0.05437	0.0473, 0.9683, 0.9681
					Coarse Tree	0.07598	0.0697, 0.9297, 0.9281
					Optimizable Tree	0.042066	0.0356, 0.9821, 0.9821
					Boosted Trees	0.14505	0.1453, 0.6707, 0.5150
					Bagged Trees	0.027565	0.0235, 0.9931, 0.9919
					Optimizable Ensemble	**0.020137**	**0.0168, 0.9965, 0.9959**
Yes	Yes	No	5% Outliers (50 features)	valence	Fine Tree	0.03189	0.0288, 0.9749, 0.9748
					Medium Tree	0.038019	0.0348, 0.9628, 0.9626
					Coarse Tree	0.054219	0.0505, 0.9201, 0.9184
					Optimizable Tree	0.029819	0.0266, 0.9786, 0.9786
					Boosted Trees	0.10072	0.0999, 0.6673, 0.4997
					Bagged Trees	0.018044	0.0146, 0.9946, 0.9932
					Optimizable Ensemble	**0.012644**	**0.0099, 0.9976, 0.9969**

Table 73 presents the arousal and valence prediction performances obtained by applying feature selection, with 15% of outliers.

Table 73. Prediction Performances using RECOLA's Physiological Data, after Feature Scaling, Standardization, and 15% Feature Selection

Scaled?	Standardized (z-score)?	Has NaNs?	Feature Selection	Prediction	Regressor	Validation RMSE	Testing RMSE, PCC, CCC
Yes	Yes	No	15% (103 features)	arousal	Fine Tree	0.047148	0.0414, 0.9758, 0.9757
					Medium Tree	0.055274	0.0494, 0.9653, 0.9652
					Coarse Tree	0.076491	0.0697, 0.9295, 0.9278
					Optimizable Tree	0.045901	0.0397, 0.9778, 0.9778
					Boosted Trees	0.14364	0.1440, 0.6821, 0.5237
					Bagged Trees	**0.029542**	**0.0248, 0.9923, 0.9910**
					Optimizable Ensemble	0.054835	0.0533, 0.9597, 0.9574
Yes	Yes	No	15% (103 features)	valence	Fine Tree	0.026529	0.0232, 0.9837, 0.9837
					Medium Tree	0.032024	0.0279, 0.9763, 0.9762
					Coarse Tree	0.046567	0.0420, 0.9455, 0.9447
					Optimizable Tree	0.0255	0.0221, 0.9852, 0.9851
					Boosted Trees	0.09946	0.0988, 0.6752, 0.5162
					Bagged Trees	0.015453	0.0132, 0.9954, 0.9945
					Optimizable Ensemble	**0.010682**	**0.0083, 0.9985, 0.9978**

APPENDIX C: GROUPS FOR VIRTUAL REALITY (VR) EXPERIMENTS

This appendix includes tables summarizing the experimental groups for the data collection process in VR. Table 74 details the tasks carried out by the participants of the first group of experiments. The Q#'s in the tables refer to specific built-in questions in the virtual environment which need to be answered by the participants to continue their level, while the PAUSEs are where the experimenter asked the four questions to the participants. The difference in the groups refers to which level they were exposed to during the immersion. Participants were randomly assigned to one of the three groups.

Table 74. Group 1 of VR Immersion Experiments

Virtual Environment	RESTAURANT				APARTMENT		BUS	
Level	**R1**	**R2**	**R3**	**R4**	**A1**	**A3**	**B1**	**B4**
Tasks	Instructions	Instructions	Instructions	**PAUSE**	Instructions	Meet friend	Instructions	Map (route 3, bus 47)
	PAUSE	**PAUSE**	**PAUSE**	Table	**PAUSE**	*Q11, Q12*	Map (rt.1, bus 3)	Go on bus
	Table	Table	Table	Take order	Go to couch	**PAUSE**	Go on bus	**PAUSE**
	Take order	Take order	Take order	**PAUSE**	Walk to phone	Go to kitchen	**PAUSE**	Stay on bus (30s)

Virtual Environment	RESTAURANT				APARTMENT		BUS	
Level	**R1**	**R2**	**R3**	**R4**	**A1**	**A3**	**B1**	**B4**
	PAUSE	**PAUSE**	**PAUSE**	Bar	**PAUSE**	**PAUSE**	Stay on bus (30s)	**PAUSE**
	Bar	Bar	Bar		Pick up phone	*Q13*	**PAUSE**	Stay on bus (30s)
					Q1	Take salad	Stay on bus	**PAUSE**
					PAUSE	**PAUSE**	Stop: *Park Avenue*	Stay on bus (30s)
					Go to kitchen	Go to living room	**PAUSE**	**PAUSE**
					Open fridge	**PAUSE**		Stay on bus (30s)
					Q2	Put salad on table		**PAUSE**
					PAUSE	*Q14*		Stop: *Bay Street*
					Go to microwave	**PAUSE**		Go in store
					Open microwave			Get bread
					PAUSE			*Q4, Q5, Q6, Q7, Q8*
					Go to entrance			**PAUSE**
					PAUSE			Cross the street
					Open the door			Map (route 2, bus 85)
								Go on bus
								PAUSE
								Stay on bus (30s)
								PAUSE
								Stop: *Midland*

Virtual Environment	RESTAURANT				APARTMENT		BUS	
Level	**R1**	**R2**	**R3**	**R4**	**A1**	**A3**	**B1**	**B4**
Immersion Time	2:30	2:30	2:30	2:30	5:00	5:00	5:20	16:00
PAUSES	2	2	2	2	6	5	3	8

Table 75 details the tasks carried out by the participants of the second group of experiments.

Table 75. Group 2 of VR Immersion Experiments

Virtual Environment	RESTAURANT				APARTMENT		BUS	
Level	**R1**	**R2**	**R3**	**R4**	**A1**	**A4**	**B1**	**B3**
Tasks	Instructions	Instructions	Instructions	**PAUSE**	Instructions	Meet friend	Instructions	Map (route 1, bus 12)
	PAUSE	**PAUSE**	**PAUSE**	Table	**PAUSE**	*Q15, Q16, Q17, Q18*	Map (rt.1, bus 3)	Go on bus
	Table	Table	Table	Take order	Go to couch	**PAUSE**	Go on bus	**PAUSE**
	Take order	Take order	Take order	**PAUSE**	Walk to phone	Go open fridge	**PAUSE**	Stay on bus (30s)
	PAUSE	**PAUSE**	**PAUSE**	Bar	**PAUSE**	**PAUSE**	Stay on bus (30s)	**PAUSE**
	Bar	Bar	Bar		Pick up phone	Go to living room	**PAUSE**	*Q2*
					Q1	**PAUSE**	Stay on bus	Stop: *King Road*
					PAUSE	Take money	Stop: *Park Avenue*	Map (route 4, bus 42)
					Go to kitchen	Go to fridge	**PAUSE**	**PAUSE**
					Open fridge	**PAUSE**		Go on bus

Virtual Environment	RESTAURANT				APARTMENT		BUS	
Level	**R1**	**R2**	**R3**	**R4**	**A1**	**A4**	**B1**	**B3**
					Q2	Choose coupon		**PAUSE**
					PAUSE	Go to phone		Stay on bus (30s)
					Go to microwave	**PAUSE**		**PAUSE**
					Open microwave	*Q19*		*Q3*
					PAUSE	Go to living room		**PAUSE**
					Go to entrance	**PAUSE**		Stop: *Main Street*
					PAUSE	*Q20*		
					Open the door	Go to entrance		
						PAUSE		
						Open the door		
						Q21		
Immersion Time	2:30	2:30	2:30	2:30	5:00	8:00	5:20	11:30
PAUSES	2	2	2	2	6	7	3	6

Table 76 details the tasks carried out by the participants of the third group of experiments.

Table 76. Group 3 of VR Immersion Experiments

Virtual Environment	RESTAURANT				APARTMENT		BUS	
Level	**R1**	**R2**	**R3**	**R4**	**A1**	**A2**	**B1**	**B2**
Tasks	Instructions	Instructions	Instructions	**PAUSE**	Instructions	Open fridge	Instructions	Map (route 3, bus 12)

Virtual Environment	RESTAURANT				APARTMENT		BUS	
Level	**R1**	**R2**	**R3**	**R4**	**A1**	**A2**	**B1**	**B2**
	PAUSE	**PAUSE**	**PAUSE**	Table	**PAUSE**	*Q3*	Map (rt.1, bus 3)	*Q1*
	Table	Table	Table	Take order	Go to couch	**PAUSE**	Go on bus	Go on bus
	Take order	Take order	Take order	**PAUSE**	Walk to phone	Grab stir-fry	**PAUSE**	**PAUSE**
	PAUSE	**PAUSE**	**PAUSE**	Bar	**PAUSE**	Put in microwave	Stay on bus (30s)	Stay on bus (30s)
	Bar	Bar	Bar		Pick up phone	*Q4*	**PAUSE**	**PAUSE**
					Q1	**PAUSE**	Stay on bus	Stay on bus (30s)
					PAUSE	Go to sink	Stop: *Park Avenue*	**PAUSE**
					Go to kitchen	*Q5*	**PAUSE**	Stay on bus (30s)
					Open fridge	**PAUSE**		**PAUSE**
					Q2	Go to bathroom		Stay on bus
					PAUSE	**PAUSE**		Stop: *Main Street*
					Go to microwave	Bring bucket to sink		Map (route 3, bus 12)
					Open microwave	**PAUSE**		*Q1*
					PAUSE	*Q6*		Go on bus
					Go to entrance	Go to phone		**PAUSE**
					PAUSE	*Q7*		
					Open the door	**PAUSE**		
						Go open front door		

Virtual Environment	RESTAURANT				APARTMENT		BUS	
Level	**R1**	**R2**	**R3**	**R4**	**A1**	**A2**	**B1**	**B2**
						Go to kitchen		
						PAUSE		
						Take bill		
						Q8, Q9		
						PAUSE		
						Go to microwave		
						Q10		
Immersion Time	2:30	2:30	2:30	2:30	5:00	12:00	5:20	8:00
PAUSES	2	2	2	2	6	8	3	5

APPENDIX D: PREDICTION PERFORMANCES OF OTHER REGRESSORS

Table 77 summarizes the prediction performances of other machine learning algorithms (i.e., regressors) that were trained and tested to predict arousal values using the combined set of physiological and visual data from the VR experiments.

Table 77. Prediction Performances of Other Regressors at Predicting Arousal Values on Combined Physiological and Visual Features from VR Immersions

Prediction	Data Nature	Eye/Facial Features	Regressor	Validation RMSE	Testing RMSE, PCC, CCC
Arousal	Physiological and Eye/Facial Features	33	Linear	0.66112	0.6196, 2.3996e-04, 1.9681e-04
			Interactions Linear	7.6394	9.0898, -0.0246, -0.0029
			Robust Linear	0.68000	0.6253, -0.0086, -0.0072
			Stepwise Linear	0.99704	0.7889, 0.0348, 0.0339
			Fine Tree	0.70836	0.8118, -0.2116, -0.2077
			Medium Tree	0.66016	0.7538, -0.1939, -0.1884
			Coarse Tree	0.60133	0.6496, -0.1932, -0.1534
			Optimizable Tree	0.58161	0.5579, -0.0150, -0.0063
			Linear SVM	0.63802	0.6315, -0.0317, -0.0269
			Quadratic SVM	0.86396	0.8094, -0.1484, -0.1479
			Cubic SVM	2.1084	1.0597, -0.1258, -0.1141
			Fine Gaussian SVM	0.59969	0.5390, 0.1214, 0.0145
			Medium Gaussian SVM	0.61819	0.6256, -0.0972, -0.0756
			Coarse Gaussian SVM	0.62002	0.5838, 0.0185, 0.0081

Prediction	Data Nature	Eye/Facial Features	Regressor	Validation RMSE	Testing RMSE, PCC, CCC
			Optimizable SVM	0.58161	0.5427, -2.9794e-17, -6.088e-33
			Efficient Linear Least Squares	0.64076	0.6181, -3.4454e-04, -2.8228e-04
			Efficient Linear SVM	0.67426	0.6356, -0.0314, -0.0272
			Optimizable Efficient Linear	0.61336	0.6072, -0.0195, -0.0151
			Rational Quadratic GPR	0.59251	0.5593, -0.0985, -0.0327
			Squared Exponential GPR	0.59251	0.5593, -0.0985, -0.0327
			Matern 5/2 GPR	0.59188	0.5579, -0.0869, -0.0284
			Exponential GPR	0.59006	0.5554, -0.0735, -0.0226
			Optimizable GPR	0.58064	0.5569, -0.0584, -0.0201
			SVM Kernel	0.62776	0.6306, -0.0669, -0.0549
			Least Squares Regression Kernel	0.59520	0.6138, -0.2127, -0.1348
			Optimizable Kernel	0.58320	0.5984, -0.1495, -0.0914
			Boosted Trees	0.62149	0.6418, -0.1660, -0.1276
			Bagged Trees	0.59868	0.6155, -0.0713, -0.0531
			Optimizable Ensemble	0.57499	0.5706, -0.0703, -0.0331
			Narrow Neural Network	1.4252	2.1297, -0.3014, -0.1517
			Medium Neural Network	1.1565	1.2447, -0.1025, -0.0814
			Wide Neural Network	1.1262	1.1659, -0.0936, -0.0774
			Bilayered Neural Network	1.1406	1.6332, 0.1257, 0.0755
			Trilayered Neural Network	1.0470	1.3777, -0.1329, -0.0977
			Optimizable Neural Network	0.58499	0.5512, NaN, NaN

APPENDIX E: COMPUTATIONAL LOAD AND COMPLEXITY OF MACHINE LEARNING MODELS

The experimental results in this thesis support the effectiveness of the proposed models at predicting affective states from multimodal data collected during VR immersion, validating their potential to enhance adaptive interventions in cognitive rehabilitation. This appendix provides details regarding the computational load and/or complexity of the implemented machine learning models. Such aspects indicate the heaviness of the proposed solution(s) in order to help evaluate its practical feasibility. Table 78 details some of these aspects for the implemented classical regressors. Table 79 and Table 80 detail aspects for the implemented deep regressors.

Table 78. Computational Load and Complexity Classical Regressors

Prediction	Data Type	Training Data	Processing	Number of Features	Regressor	Model Size	Prediction Speed (predictions per second)	Training Time (seconds)	Training Time (hours)
Arousal	Physiological	Predesigned Features	Removal of NaNs, min-max normalization, z-score standardization, and 5% outliers feature selection	50 (out of 118)	Optimizable Ensemble	312 MB	20000	10653	3
	Visual	Full-Face Predesigned Features	Min-max normalization, and z-score standardization	40		452 MB	2700	45849	12.7
	Acoustic	Predesigned Features	Min-max normalization and z-score standardization	130		833 MB	12000	18902	5.3
	Multimodal	Various (full-face)	various	220		405 MB	20000	44279	12.3
		Various (half-face)	various	189		98 MB	77000	16171	4.5
Valence	Physiological	Predesigned Features	Removal of NaNs, min-max normalization, z-score standardization, and 15% outliers feature selection	103 (out of 118)		710 MB	4900	27087	7.5
	Visual	Predesigned Features	Min-max normalization, and z-score standardization	40		74 MB	55000	13669	3.8
	Acoustic	Predesigned Features	Min-max normalization and z-score standardization	130		413 MB	10000	18113	5
	Multimodal	Various (full-face)	various	273		16 MB	150000	11913	3.3
		Various (half-face)	various	242		643 MB	110000	33240	9.2

Table 79. Computational Load and Complexity Deep Regressors

Prediction	Data Type	Training Data	Neural Network	Network Layers	Number of Learnables	Training Epochs	Iterations per Epoch	Total Iterations	Validation Frequency (iterations)	Training Time (seconds)	Training Time (hours)
Arousal	Physiological	ECG and EDA sequences	Sequence-to-label RNN	4	128.2K	30	9440	283200	4720	7576	2.1
	Visual	Full-face images	MobileNet-v2	153	2.2M	30	9440	283200	4720	99543	27.7
		Full-face images	ResNet-18	70	11.1M	30	9440	283200	4720	187617	52.1
		Half-face images	MobileNet-v2	153	2.2M	30	9440	283200	4720	107595	29.9
		Half-face images	ResNet-18	70	11.1M	30	9440	283200	4720	64352	17.9
Valence	Physiological	ECG and EDA sequences	Sequence-to-label RNN	4	128.2K	30	9440	283200	4720	7260	2
	Visual	Full-face images	MobileNet-v2	153	2.2M	30	9440	283200	4720	84028	23.3
		Full-face images	ResNet-18	70	11.1M	30	9440	283200	4720	60446	16.8
		Half-face images	MobileNet-v2	153	2.2M	30	9440	283200	4720	107557	29.9
		Half-face images	ResNet-18	70	11.1M	30	9440	283200	4720	64951	18

Table 80. Size of Implemented Deep Regressors

Neural Network	Depth (layers)	Parameter Memory	Parameters/Learnables (Millions)	Image Input Size	Input Value Range	Input Layer Normalization	Connections
MobileNet-v2	153	14 MB	2.2	280×280 (full faces) **OR** 140×280 (half faces)	[0, 255]	z-score	162
ResNet-18	70	45 MB	11.1				77

BIBLIOGRAPHY

1. Schmidt, P.; Reiss, A.; Dürichen, R.; Laerhoven, K.V. Wearable-Based Affect Recognition—A Review. *Sensors* **2019**, *19*, 4079.

2. Ayoub, I. Multimodal Affective Computing Using Temporal Convolutional Neural Network and Deep Convolutional Neural Networks. Master's Thesis, University of Ottawa, Ottawa, ON, Canada, 2019. Available online: https://ruor.uottawa.ca/bitstream/10393/39337/1/Ayoub_Issa_2019_Thesis.pdf (accessed on 28 May 2020).

3. Corneanu, C.A.; Simon, M.O.; Cohn, J.F.; Guerrero, S.E. Survey on RGB, 3D, Thermal, and Multimodal Approaches for Facial Expression Recognition: History, Trends, and Affect-Related Applications. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *38*, pp. 1548-1568.

4. Al Osman, H.; Falk, T.H. Multimodal affect recognition: Current approaches and challenges. *Emot. Atten. Recognit. Based Biol. Signals Images* **2017**, *8*, pp. 59-86.

5. VanHorn, K.C.; Zinn, M.; Cobanoglu, M.C. Deep Learning Development Environment in Virtual Reality. pp. 1-10, 2019.

6. Russell, J. *Affective Space Is Bipolar*; American Psychological Association: Washington, DC, USA, 1979.

7. Ringeval, F.; Sonderegger, A.; Sauer, J.; Lalanne, D. Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions. In Proceedings of the 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), Shanghai, China, 22-26 April 2013; pp. 1-8.

8. Sun, B.; Sun, B.; Li, L.; Zhou, G.; Wu, X.; He, J.; Yu, L.; Li, D.; Wei, Q. Combining multimodal features within a fusion network for emotion recognition in the wild. In Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, Seattle, WA, USA, 9-13 November 2015; pp. 497-502.

9. Sun, B.; Li, L.; Zhou, G.; He, J. Facial expression recognition in the wild based on multimodal texture features. *J. Electron. Imaging* **2016**, *25*, 061407.

10. Dhall, A.; Goecke, R.; Ghosh, S.; Joshi, J.; Hoey, J.; Gedeon, T. From individual to group-level emotion recognition: EmotiW 5.0. In Proceedings of the 19th ACM International Conference on Multimodal Interaction, Glasgow, UK, 13-17 November 2017; pp. 524-528.

11. Alpaydin, E. Introduction to Machine Learning. Massachusetts Institute of Technology (MIT) Press: Cambridge, MA, 2020.

12. Jordan, M. I.; Mitchell, T. M. Machine learning: Trends, perspectives, and prospects. *Science* **2015**, *349*, 6245, pp. 255-260.

13. Help Center. Help Center for MATLAB, Simulink, and Other MathWorks Products. Available online: https://www.mathworks.com/help/ (accessed on 2 September 2022).

14. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 7553, pp. 436-444.

15. Hou, L.; Samaras, D.; Kurc, T. M.; Gao, Y.; Davis, J. E.; Saltz, J. H. Patch-based Convolutional Neural Network for Whole Slide Tissue Image Classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016; pp. 2424-2433.

16. Von Zitzewitz, G. Survey of neural networks in autonomous driving. July 2017.

17. Vardhana, M.; Arunkumar, N.; Lasrado, S.; Abdulhay, E.; Ramirez-Gonzalez, G. Convolutional neural network for bio-medical image segmentation with hardware acceleration. *Cognitive Systems Research* **2018**, *50*, pp. 10-14.

18. Valstar, M.; Gratch, J.; Schuller, B.; Ringeval, F.; Lalanne, D.; Torres, M.T.; Scherer, S.; Stratou, G.; Cowie, R.; Pantic, M. AVEC 2016—Depression, Mood, and Emotion Recognition Workshop and Challenge. In Proceedings of the AVEC'16, Amsterdam, The Netherlands, 16 October 2016; ACM: New York, NY, USA, 2016; pp. 3-10.

19. Gunes, H.; Schuller, B. Categorical and dimensional affect analysis in continuous input: Current trends and future directions. Image Vis. Comput. 2013, 31, pp. 120-136.

20. Almaev, T. R.; Valstar, M. F. Local Gabor Binary Patterns from Three Orthogonal Planes for Automatic Facial Expressions Recognition. In Proceedings of ACII, Geneva, Switzerland, 2013; IEEE Computer Society, 356-361.

21. Xiong, X.; De la Torre, F. Supervised descent method and its applications to face alignment. In Proceedings of CVPR, Portland (OR), USA, 2013; IEEE, 532-539.

22. Ekman, P.; Friesen, W.V. Facial action coding system: A technique for the measurement of facial movement; Consulting Psychologists Press: Palo Alto, CA, 1978.

23. Ringeval, F.; Eyben, F.; Kroupi, E.; Yuce, A.; Thiran, J.P.; Ebrahimi, T.; Lalanne, D.; Schuller, B. Prediction of Asynchronous Dimensional Emotion Ratings from Audiovisual and Physiological Data. *Pattern Recognition Letters* **2015**, 66, 22-30.

24. Recola Data set. Available online: https://diuf.unifr.ch/main/diva/recola/ (accessed on 28 May 2022).

25. Ringeval, F.; Schuller, B; Valstar, M; Jaiswal, S; Marchi, E; Lalanne, D; Cowie, R; Pantic, M. AV+EC 2015 – The First Affect Recognition Challenge Bridging Across Audio, Video, and Physiological Data. In Proceedings of AVEC'15, Brisbane, Australia, 2015; ACM, New York, NY, USA, 3–8.

26. Ringeval, F.; Schuller, B.; Valstar, M.; Cowie, R.; Kaya, H.; Schmitt, M.; Amiriparian, S.; Cummins, N.; Lalanne, D.; Michaud, A.; et al. AVEC 2018 Workshop and Challenge: Bipolar Disorder and Cross-Cultural Affect Recognition. In Proceedings of the AVEC'18, Seoul, Republic of Korea, 22 October 2018; ACM, New York, NY, USA, 2018.

27. Amirian, M.; Kächele, M.; Thiam, P.; Kessler, V.; Schwenker, F. Continuous Multimodal Human Affect Estimation using Echo State Networks. In Proceedings

of the 6th ACM International Workshop on Audio/Visual Emotion Challenge (AVEC'16), Amsterdam, The Netherlands, 16 October 2016; pp. 67-74.

28. Tzirakis, P.; Zafeiriou, S.; Schuller, B.W. End2You—The Imperial Toolkit for Multimodal Profiling by End-to-End Learning. *arXiv* **2018**, arXiv:1802.01115.

29. Brady, K.; Gwon, Y.; Khorrami, P.; Godoy, E.; Campbell, W.; Dagli, C.; Huang, T.S. Multi-Modal Audio, Video and Physiological Sensor Learning for Continuous Emotion Prediction. In Proceedings of the AVEC'16, Amsterdam, The Netherlands, 16 October 2016; pp. 97-104.

30. Han, J.; Zhang, Z.; Ren, Z.; Schuller, B. Implicit Fusion by Joint Audiovisual Training for Emotion Recognition in Mono Modality. In Proceedings of the 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12-17 May 2019; pp. 5861-5865.

31. Weber, R.; Barrielle, V.; Soladié, C.; Séguier, R. High-Level Geometry-based Features of Video Modality for Emotion Prediction. In Proceedings of the AVEC'16, Amsterdam, The Netherlands, 16 October 2016; pp. 51-58.

32. Albadawy, E.; Kim, Y. Joint Discrete and Continuous Emotion Prediction Using Ensemble and End-to-End Approaches. In Proceedings of the 20th ACM International Conference on Multimodal Interaction (ICMI '18), 2018; ACM, New York, NY, USA, 366-375.

33. Khorrami, P.; Paine, T.L.; Brady, K.; Dagli, C.; Huang, T.S. How Deep Neural Networks Can Improve Emotion Recognition on Video Data. In Proceedings of the IEEE International Conference on Image Processing (ICIP), Phoenix, AZ, USA, 25-28 September 2016; pp. 619-623.

34. Povolny, F.; Matejka, P.; Hradis, M.; Popková, A.; Otrusina, L.; Smrz, P.; Wood, I.; Robin, C.; Lamel, L. Multimodal emotion recognition for AVEC 2016 challenge. In Proceedings of the AVEC'16, Amsterdam, The Netherlands, 16 October 2016; pp. 75-82.

35. Somandepalli, K.; Gupta, R.; Nasir, M.; Booth, B.M.; Lee, S.; Narayanan, S.S. Online affect tracking with multimodal kalman filters. In Proceedings of the AVEC'16, Amsterdam, The Netherlands, 16 October 2016; pp. 59-66.

36. Nicolaou, M.A.; Gunes, H.; Pantic, M. Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space. *IEEE Transactions on Affective Computing.* **2011**, *2*, pp. 92-105.

37. Gunes, H.; Piccardi, M.; Pantic, M. From the Lab to the Real World: Affect Recognition Using Multiple Cues and Modalities. In *Affective Computing*; I-Tech Education and Publishing: Mumbai, India, 2008; pp. 185-218.

38. Ringeval, F.; Eyben, F.; Kroupi, E.; Yuce, A.; Thiran, J.P.; Ebrahimi, T.; Lalanne, D.; Schuller, B. Pattern Recognition Letters Prediction of Asynchronous Dimensional Emotion Ratings from Audiovisual and Physiological Data. *Pattern Recognit. Lett.* **2015**, *66*, pp. 22-30.

39. Chen, S.; Jin, Q. Multi-modal conditional attention fusion for dimensional emotion prediction. In Proceedings of the 24th ACM International Conference on Multimedia, Amsterdam, The Netherlands; 15-19 October 2016; pp. 571-575.

40. Tzirakis, P.; Trigeorgis, G.; Nicolaou, M.A.; Schuller, B.; Zafeiriou, S. End-to-End Multimodal Emotion Recognition using Deep Neural Networks. *IEEE J. Sel. Top. Signal Process.* **2017**, *11*, pp. 1301-1309.

41. Huang, Y.; Lu, H. Deep learning driven hypergraph representation for image-based emotion recognition. In Proceedings of the 18th ACM International Conference on Multimodal Interaction (ICMI 2016), Tokyo Japan; 12-16 November 2016; pp. 243-247.

42. Kahou, S.E.; Michalski, V.; Konda, K.; Memisevic, R.; Pal, C. Recurrent neural networks for emotion recognition in video. In Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, Seattle, WA, USA, 9-13 November 2015; pp. 467-474.

43. Bogie, B.J.M.; Noël, C.; Gu, F.; Nadeau, S.; Shvetz, C.; Khan, H.; Rivard, M.-C.; Bouchard, S.; Lepage, M.; Guimond, S. Using virtual reality to improve verbal episodic memory in schizophrenia: A proof-of-concept trial. *Schizophrenia Research: Cognition* **2024**, *36*, 100305.

44. Bekele, E.; Bian, D.; Peterman, J.; Park, S.; Sarkar, N. Design of a Virtual Reality System for Affect Analysis in Facial Expressions (VR-SAAFE); Application to Schizophrenia. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* **2017**, *25*, *6*, pp. 739-749.

45. Dawson, M. E.; Nuechterlein, K. H. Psychophysiological dysfunctions in the developmental course of schizophrenic disorders. *Schizophrenia Bulletin* **1984**, *10*, *2*, pp. 204-232.

46. Ikezawa, S.; Corbera, S.; Liu, J.; Wexler, B. E. Empathy in electrodermal responsive and nonresponsive patients with schizophrenia. *Schizophrenia Research* **2012**, *142*, *1*, pp. 71-76.

47. Kring, A. M.; Kerr, S. L.; Smith, D. A.; Neale, J. M. Flat affect in schizophrenia does not reflect diminished subjective experience of emotion. *Journal of Abnormal Psychology* **1993**, *102*, *4*, pp. 507-517.

48. Kring, A. M.; Neale, J. M. Do schizophrenic patients show a disjunctive relationship among expressive, experiential, and psychophysiological components of emotion?. *Journal of Abnormal Psychology* **1996**, *105*, *2*, pp. 249-257.

49. Herbener, E. S.; Song, W.; Khine, T. T.; Sweeney, J. A. What aspects of emotional functioning are impaired in schizophrenia?. *Schizophrenia Research* **2008**, *98*, *1*, pp. 239-246.

50. Hempel, R. J. et al. Physiological responsivity to emotional pictures in schizophrenia. *Journal of Psychiatric Research* **2005**, *39*, *5*, pp. 509-518.

51. Cacioppo, J. T.; Tassinary, L. G.; Berntson, G. Handbook of Psychophysiology. Cambridge, U.K.: Cambridge University Press, 2007.

52. Liu, C.; Conn, K.; Sarkar, N.; Stone, W. Online affect detection and robot behavior adaptation for intervention of children with autism. *IEEE Transactions on Robotics* **2008**, *24*, *4*, pp. 883-896.

53. Welch, K. C. et al. An affect-sensitive social interaction paradigm utilizing virtual reality environments for autism intervention. In Proceedings of the International Conference on Human-Computer Interaction, 2009; pp. 703-712.

54. Zhang, W.; Shu, L.; Xu, X.; Liao, D. Affective Virtual Reality System (AVRS): Design and Ratings of Affective VR Scenes. In Proceedings of the International Conference on Virtual Reality and Visualization (ICVRV), 2017.

55. White, S. E.; Dittrich, J. E.; Lang, J. R. The effects of group decision-making process and problem situation complexity on implementation attempts. *Administrative Science Quarterly* **1980**, *25*, *3*, pp. 428-440.

56. Joyal, C. C.; Jacob, L.; Cigna, M.; Guay, J.; Renaud, P. Virtual faces expressing emotions: an initial concomitant and construct validity study. *Frontiers in Human Neuroscience* **2014**, *8*, *787*, pp. 1-6.

57. Dyck, M.; Winbeck, M.; Leiberg, S.; Chen, Y.; Gur, R. C.; Mathiak, K. Recognition Profile of Emotions in Natural and Virtual Faces. *PLoS ONE* **2008**, *3*, *11*, *e3628*, pp. 1-8.

58. Benlamine, M.S.; Dufresne, A.; Beauchamp, M.H. et al. BARGAIN: behavioral affective rule-based games adaptation interface–towards emotionally intelligent games: application on a virtual reality environment for socio-moral development. *User Modeling and User-Adapted Interaction* **2021**, *31*, pp. 287-321.

59. Granato, M.; Gadia, D.; Maggiorini, D. et al. An empirical study of players' emotions in VR racing games based on a dataset of physiological data. *Multimedia Tools and Applications* **2020**, *79*, pp. 33657-33686.

60. Warp, R.; Zhu, M.; Kiprijanovska, I.; Wiesler, J.; Stafford S.; Mavridou, I. Validating the effects of immersion and spatial audio using novel continuous biometric sensor measures for Virtual Reality. In Proceedings of IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct), Singapore, Singapore, 2022; pp. 262-265.

61. Breiman, L. Random Forests. *Machine Learning* **2001**, *45*, pp. 5-32.

62. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. Preprint, submitted 10 December, 2015. https://arxiv.org/abs/1512.03385.

63. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.-C. MobileNetV2: Inverted Residuals and Linear Bottlenecks. In proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 2018; pp. 4510-4520.

64. Joudeh, I. O.; Cretu, A.-M.; Wallace, B.; Goubran, R.; Alkhalid, A.; Allegue-Martinez, M; Knoefel, F. WiFi Channel State Information-Based Recognition of Sitting-Down and Standing-Up Activities. In Proceedings of the 2019 IEEE International Symposium on Medical Measurements and Applications (MeMeA), Istanbul, Turkey, 2019; pp. 1-6.

65. Viola, P.; Jones, M. Rapid Object Detection using a Boosted Cascade of Simple Features. In Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2001), Kauai, HI, USA, 8-14 December 2001; pp. 511-518.

66. Schizophrenia. World Health Organization. Available online: https://www.who.int/news-room/fact-sheets/detail/schizophrenia (accessed on 3 March 2024).

67. Vive Pro Eye Specs & User Guide. Available online: https://developer.vive.com/resources/hardware-guides/vive-pro-eye-specs-user-guide/ (accessed on 3 March 2024).

68. Everitt, B. The Cambridge Dictionary of Statistics. Cambridge, U.K.: Cambridge University Press, 1998.

69. Ding, C.; Peng, H. Minimum redundancy feature selection from microarray gene expression data. *Journal of Bioinformatics and Computational Biology* **2005**, *3*, *2*, pp. 185-205.

70. Yang, W.; Wang, K.; Zuo, W. Neighborhood Component Feature Selection for High-Dimensional Data. *Journal of Computers* **2012**. *7*, *1*.

71. Kononenko, I.; Šimec, E.; Robnik-Šikonja, M. Overcoming the Myopia of Inductive Learning Algorithms with RELIEFF. *Applied Intelligence* **1997**, *7*, pp. 39-55.

72. Robnik-Sikonja, M.; Kononenko, I. An adaptation of Relief for attribute estimation in regression. In Proceedings of the International Conference on Machine Learning, 1997.

73. Robnik-Sikonja, M.; Kononenko, I. Theoretical and empirical analysis of ReliefF and RReliefF. *Machine Learning* **2003**, *53*, pp. 23-69.

74. Alwosheel, A.; van Cranenburgh, S.; Chorus, C.G. Is your dataset big enough? Sample size requirements when using artificial neural networks for discrete choice analysis. *Journal of Choice Modelling* **2018**, *28*, pp. 167-182.

75. Haykin, S. 2009. *Neural Networks and Learning Machines (3rd ed.)*. Upper Saddle River: Pearson Education.

76. Abu-Mostafa, Y.S. Hints. *Neural Computation* **1995**, *7*, *4*, pp. 639-71.

77. Baum, E.B.; Haussler, D. What Size Net Gives Valid Generalization?. Neural Computation **1989**, *1*, *1*, pp. 151-160.

78. Joudeh, I.O.; Cretu, A.-M.; Guimond, S.; Bouchard, S. Prediction of Emotional Measures via Electrodermal Activity (EDA) and Electrocardiogram (ECG). *Eng. Proc.* **2022**, *27*, 47.

79. Joudeh, I.O.; Cretu, A.-M.; Bouchard, S.; Guimond, S. Prediction of Continuous Emotional Measures through Physiological and Visual Data. *Sensors* **2023**, *23*, 5613.

80. Joudeh, I.O.; Cretu, A.-M.; Bouchard, S.; Guimond, S. Prediction of Emotional States from Partial Facial Features for Virtual Reality Applications. *Annual Review of Cybertherapy and Telemedicine* **2023**, *21*.

81. Joudeh, I.O.; Cretu, A.; Bouchard, S. Optimizable Ensemble Regression for Arousal and Valence Predictions from Visual Features. In Proceedings of the 10th International Electronic Conference on Sensors and Applications, 15-30 November 2023, MDPI: Basel, Switzerland.

82. Joudeh, I.O.; Cretu, A.-M.; Bouchard, S. Predicting the Arousal and Valence Values of Emotional States Using Learned, Predesigned, and Deep Visual Features. *Sensors* **2024**, *24*, 4398.

83. Rahman, M. M.; Rivolta, M. W.; Badilini, F.; Sassi, R. A Systematic Survey of Data Augmentation of ECG Signals for AI Applications. *Sensors* **2023**. *23*, *11*, 5237.