Summer 8-2022

Emotion Detection Using an Ensemble Model Trained with Physiological Signals and Inferred Arousal-Valence States

Matthew Nathanael Gray
Old Dominion University, mgray564@gmail.com

EMOTION DETECTION USING AN ENSEMBLE MODEL TRAINED WITH PHYSIOLOGICAL SIGNALS AND INFERRED AROUSAL-VALENCE STATES

by

Matthew Nathanael Gray
B.S. April 2016, University of Alabama at Birmingham

A Thesis Submitted to the Faculty of
Old Dominion University in Partial Fulfillment of the
Requirements for the Degree of

MASTER OF SCIENCE

ELECTRICAL AND COMPUTER ENGINEERING

OLD DOMINION UNIVERSITY
August 2022

Approved by:

Jiang Li (Director)

Krzysztof Rechowicz (Member)

Sampath Jayarathna (Member)

Chung-Hao Chen (Member)

ABSTRACT

EMOTION DETECTION USING AN ENSEMBLE MODEL TRAINED WITH PHYSIOLOGICAL SIGNALS AND INFERRED AROUSAL-VALENCE STATES

Matthew Nathanael Gray
Old Dominion University, 2022
Director: Dr. Jiang Li

Affective computing is an exciting and transformative field that is gaining in popularity among psychologists, statisticians, and computer scientists. The ability of a machine to infer human emotion and mood, i.e. affective states, has the potential to greatly improve human-machine interaction in our increasingly digital world. In this work, an ensemble model methodology for detecting human emotions across multiple subjects is outlined. The Continuously Annotated Signals of Emotion (CASE) dataset, which is a dataset of physiological signals labeled with discrete emotions from video stimuli as well as subject-reported continuous emotions, arousal and valence, from the circumplex model, is used for training and testing the model [1, 2]. Blood volume pulse (BVP), galvanic skin response (GSR), and skin temperature physiological signals are windowed and used to extract 17 physiological features (13 BVP, 2 GSR, and 2 skin temperature features). These physiological features are then used along with subject-reported arousal and valence state values as inputs into regression models to create predicted arousal and valence values for each feature window. The predicted or “inferred” arousal and valence state values were then concatenated to the original 17 physiological features and used as inputs to a classification model for the final classification of emotion state into five categories, including relaxed, bored, neutral, amused, and scared. Multiple regression and classification models were tested, and the best performing model was a linear regression arousal and valence predictor followed by a hyperparameter-tuned support vector machine (SVM) classifier, achieving a five-fold cross-validation accuracy of ***98.79% ± 0.29%*** for the five-class emotion classification across subjects. Finally, an impactful real-world application in an emotional feedback household environment for enabling independent living in differently-abled people is discussed.

This thesis is dedicated to my family whose love and support has always enabled me to pursue my dreams.

ACKNOWLEDGEMENTS

I would like to sincerely thank my advisor and committee members for encouraging and guiding me throughout my thesis. Their advice and support helped me through the coding, analysis, and writing of my thesis, and they genuinely supported me throughout my progress and completion. They helped me find a good work-life balance with helpful advice on staying productive while having free time for other aspects of life.

I would also like to thank my teammates and coworkers during my graduate research assistantship at the Virginia Modeling Analysis and Simulation Center (VMASC) for always supporting me and being great mentors and friends. I would like to thank my lab mates in the Neuro-Information Retrieval and Data Science (NIRDS) lab at ODU for their sincere help and friendship as well.

NOMENCLATURE

AMIGOS	A dataset for Multimodal research of affect, personality traits, and mood on Individuals and GrOupS
AUC	Area Under the Curve
BPM	Beats Per Minute
BVP	Blood Volume Pulse
CASE	Continuously Annotated Signals of Emotion
cEMG	corrugator supercilii-Electromyogram
CNN	Convolutional Neural Network
CUDA	Compute Unified Device Architecture
DEAP	Database for Emotional Analysis using Physiological Signals
DECAF	MEG-Based Multimodal Database for DECoding AFfective Physiological Responses
DNN	Deep Neural Network
DREAMER	A Database for Emotion Recognition Through EEG and ECG Signals From Wireless Low-cost Off-the-Shelf Devices
GPU	Graphics Processing Unit
GSR	Galvanic Skin Response
ECG	Electrocardiogram
EEG	Electroencephalography
EMG	Electromyography
EOG	Electrooculogram
hEOG	horizontal-Electrooculogram
HMM	Hidden Markov Model
IBI	Inter-beat Interval
JERI	Joystick-based Emotion Reporting Interface
KNN	K-Nearest Neighbor
LDA	Linear Discriminant Analysis
LDF	Linear Discriminant Function
LSTM	Long Short-Term Memory
MAD	Median Absolute Deviation of ECG RR Intervals
MAE	Mean Absolute Error
MAHNOB-HCI	A Multimodal Database for Affect Recognition and Implicit Tagging
MAP	Maximum a Posteriori

MEG	Magnetoencephalogram
MLP	Multilayer Perceptron
MMC	Meta-multiclass
M-SVR	Multiple Output Support Vector Regression
mRMR	Minimum Redundancy Maximum Relevance
MSE	Mean Squared Error
NIR	Near Infrared
NN	Neural Network
pNN20	Proportion of Successive Differences above 20 milliseconds
pNN50	Proportion of Successive Differences above 50 milliseconds
PPG	Photoplethysmography
RECOLA	Remote Collaborative and Affective Interactions
ReLU	Rectified Linear Unit
RF	Random Forest
RFE	Recursive Feature Elimination
RMSSD	Root Mean Square of Successive Differences of Intervals
RMSE	Root Mean Squared Error
S	Area of Poincaré Ellipse
SD1	Standard Deviation Perpendicular to Line of Identity
SD2	Standard Deviation Parallel to Line of Identity
SD1/SD2	Ratio of Poincaré Standard Deviations
SDNN	Standard Deviation of ECG NN Intervals
SDSD	Standard Deviation of Successive Differences
SEMAINE	Sustained Emotionally colored Machine-human Interaction using Nonverbal Expression Dataset
SEWA	A Rich Database for Audio-Visual Emotion and Sentiment Research in the Wild
SFFS	Sequential Floating Feature Selection
SFS	Sequential Forward Selection
SMOTE	Synthetic Minority Oversampling Technique
SVM	Support Vector Machine
SVR	Support Vector Regression
tEMG	trapezius-Electromyogram
WMD-DTW	Weighted Multi-Dimensional Dynamic Time Warping
zEMG	zygomaticus major-Electromyogram

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

Figure Page

Figure Page

Figure Page

CHAPTER 1

INTRODUCTION

1.1 Background

Affective computing is the study of detecting one's emotional (affective) state through computational methods such as facial expression recognition, body language recognition, speech tone and inflection recognition, and physiological signals [3]. The ability of a computer to infer the emotional state of a human being is potentially revolutionary through applications such as improving human-machine interaction, notification of a certain emotional state for improved self-awareness and emotion regulation, assisting individuals with autism in communicating emotional states, and augmenting an individual's environment based on a certain detected emotional state. The miniaturization of electronic sensors and processors with increased computational ability has enabled the use of affordable devices which can measure and analyze signals such as heart rate, galvanic skin response, respiration rate, skin temperature, video, and sound using computationally expensive filtering and modeling techniques. The increased availability of such low-cost and miniaturized sensors has allowed the average person access to data previously only available in a lab using highly specialized equipment and software. Society is currently experiencing a data revolution where vast quantities of data are collected about people's daily lives and can be used to improve living conditions. Affective computing is one of many big data fields on the cusp of becoming mainstream through smart wearable devices. It is gaining in popularity as can be seen by the growth of the number of research papers submitted in recent years [3-10].

Many wearable devices have been created and proposed for measuring various physiological signals [11]. In Healey's dissertation in the early 2000s, multiple unique devices are proposed including a sensor embedded inside a shoe to measure GSR from the sole of the foot, a photoplethysmography (PPG) sensor for measuring blood volume pulse (BVP) at the lobe of the ear which is worn like an earring, and a respiration sensor embedded in a sports bra for measuring respiration rate through chest expansion and decompression [11]. In recent years, wearable physiological sensors, namely BVP and GSR sensors, have become commercially available in devices usually worn on the wrist that are smaller or the same size as watches. Devices such as Fitbits ®, Apple Watches ®, and Garmin ® Smartwatches are all small, affordable devices which contain embedded sensors such as BVP, GSR, and skin temperature. Research devices such as the Empatica E4 are also available which measure BVP, GSR, and skin temperature with Bluetooth capability for data acquisition and real-time use-cases. Devices will only become smaller, cheaper, and more capable as time goes on, so now is an opportune time to develop systems and processes which apply this technology to solving complex problems as well as improving everyday life.

Real-time affective computing using widely available and affordable wearable technology has the potential to improve lives, especially for those with emotional and/or cognitive differences. A typical symptom of children and adults with cognitive differences such as autism is difficulty in expressing emotions [12]. Emotions are fully experienced, but the expression of the emotion is inhibited. This leads to utilizing other methods for individuals with autism to be self-aware of and better express emotions such as affective computing. Some challenges would need to be solved such as giving the user of the affective computing system the ability to decide whether their emotion is shared or not with the outside world which is necessary for social interactions. This would be equivalent to throttling the outward expression of emotion to abide by social norms and cues. The goal of such a system would be to improve the user's life by improving their social interactions with the world through insight into their own emotions and behavior to themselves and those around them on a real-time basis. The predicted emotion from the affective computing device could be taken a step further and be used to drive feedback from a smart environment such as a house or vehicle to actively improve someone's daily experiences with the outside world [13-16].

There are several limitations to the current state-of-the-art in affective computing. The most prominent of which are low accuracies for *subject-independent* models – models that are generalizable to any user. The models in current literature with high accuracy (90% or greater) are all *subject-dependent*. Another limitation of some studies is the use of complicated sensors such as electroencephalography (EEG), electromyography (EMG), near-infrared spectroscopy (NIRs), and others that are difficult to use outside of a controlled lab setting and thus have limited real-world applications. There are also difficulties in comparing affective computing models due to the lack of standardization in emotion labels. This makes it difficult to use multiple datasets of human emotions in the same model due to the use of differing emotion labels. Another limitation is the use of ground truth emotions from subject self-reporting in some affective datasets. While this may help account for the uniqueness in emotional responses to stimuli, this also introduces noise in the form of human bias due to phenomena such as confirmation bias and the desire to follow cultural norms. A much more comprehensive discussion on the current state-of-the-art and the limitations and challenges facing the field of affective computing can be found in the second chapter, Related Work.

1.2 Proposed Work

Using the *Continuously Annotated Signals of Emotion* (CASE) dataset which includes discrete emotion labels as well as continuous arousal and valence emotion labels correlated to physiological data from thirty subjects, a novel method of preprocessing, generating features, and ensemble model generation is presented. The following research questions are investigated:

- Can a subject-independent, multi-class emotion recognition model with greater than 95% accuracy be created?
- Can this model use only non-invasive, readily available sensors such as BVP, GSR, and Skin Temperature with high accuracies?

Based on these questions, the following hypotheses are proposed and investigated in this work:

- A discrete emotion detection model using physiological sensors can be made that is generalizable to any user (subject-independent).
- A discrete emotion detection model using physiological sensors can be made with greater than 95% accuracy.
- A discrete emotion detection model can be made using easily-acquired, readily-available physiological sensors currently on the market today.

1.3 Outline of Thesis

This work covers a background of affective/emotion model types, openly available affective/emotion datasets, a literature review of the field of affective computing from its inception in the late 1980s to 2022, and brief descriptions of the regression and classification models used in this work in the Related Work section. Then a description of a novel, generalizable, and well-performing methodology for predicting emotion is given in the Materials and Methods section. Accuracy and other metrics for this methodology are given in the Results section, and discussions of potential applications and significance are given in the Discussion and Conclusions section.

CHAPTER 2

RELATED WORK

2.1 Affective Models

Some of the first models and hypotheses of emotions are from Charles Darwin and James-Lange [17]. Emotions are complex psychological phenomena derived from conscious and unconscious thoughts. In a pathological sense, emotion dysregulation due to mental illnesses such as bipolar, depression, anxiety, and borderline personality disorder can significantly affect one's life. To better understand emotions and the roles they play in our lives, many different emotion models have been developed. These emotion models vary greatly in form and function and can be categorized into two types: continuous and discrete. Continuous emotion models describe emotions in a continuum across multiple dimensions. Discrete models, on the other hand, describe emotions as separately defined phenomena. It is also worth noting that in the field of psychology, "affect" is an overarching term used to describe moods and emotions [18]. An emotion is a short-term feeling caused by a stimulus such as a thought or an experience, and a mood is a long-term state of mind lasting hours, days, or longer which do not necessarily need a stimulus [18]. The following section briefly looks at the variety of emotion model methodologies suggested by researchers, and a more detailed and comprehensive literature review is available if the reader desires to dive deeper into emotion theories in [3-10, 19]. Fig. 1 shows the varied nature and categorization of popular emotion models.

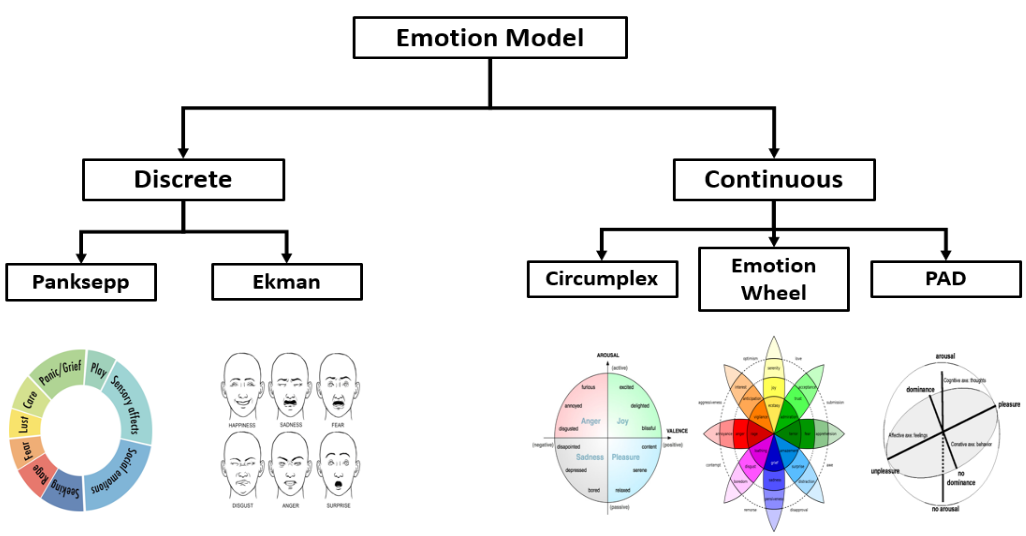


Fig. 1. Categorization of Example Emotion Models.

2.1.1 *Continuous Models*

Continuous models of affect define multiple attributes along a continuous spectrum. Each attribute is placed on a dimension, and each emotion is defined by its location in the n-dimensional space. The most popular continuous model is the two-dimensional circumplex model of affect defined by Russel which is defined by two attributes: arousal and valence [2]. There are also other continuous models such as Mehrabian's three-dimensional PAD (Pleasure-Arousal-Dominance) model [20] and Plutchik's "emotion wheel" model [21].

The circumplex model of affect defined by Russel in 1980 is a two-dimensional model of arousal and valence. Arousal is a measure of a person's excitement where low arousal depicts emotions such as boredom or relaxation and high arousal depicts amusement or anger. Valence is a measure of negativity to positivity where low valence depicts emotions such as sadness or anger and high valence depicts amusement and joy. This model can be correlated with discrete emotion labels by defining regions in the multi-dimensional space each emotion belongs. Fig. 2 below shows an example of correlating the arousal/valence circumplex model of affect to discrete emotions:

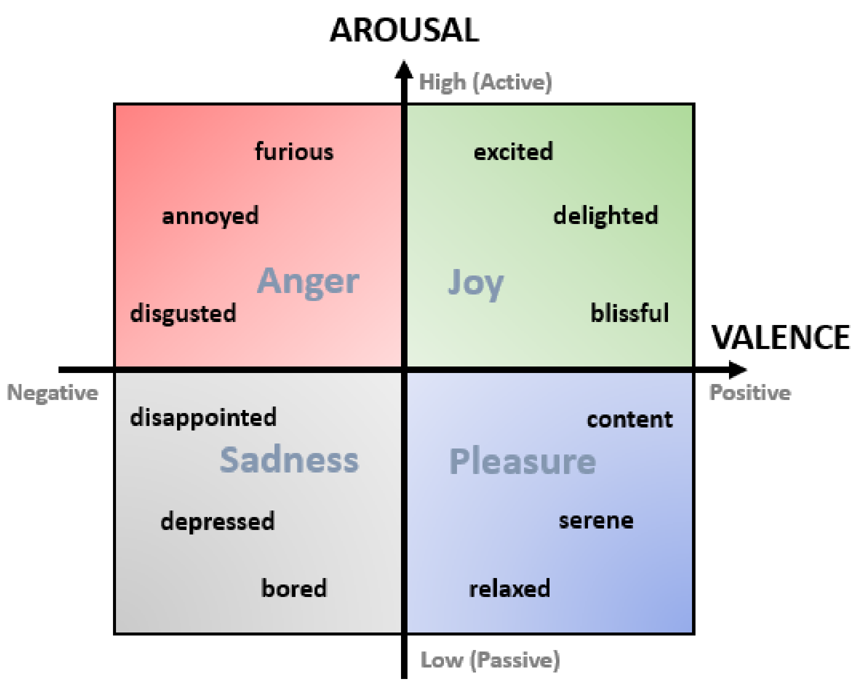


Fig. 2. Example of Arousal/Valence Space with Correlated Discrete Emotions.

The circumplex model has also been correlated with physiological responses from GSR, BVP, facial expressions, and others. One of the first studies to correlate arousal and valence with physiological responses is Winton et al. in 1984 [22]. They proved that arousal and valence responses from emotional stimuli in the form of pictures can be linearly correlated with changes in physiological features such as GSR amplitude and heart rate [22]. The two dimensions of the circumplex model (*arousal* and *valence*) are preferred over traditional discrete emotion labels such as *joy*, *anger*, and *sadness* since it better captures the time and intensity varying nature of emotions [1].

Other continuous emotion descriptors exist such as dominance, liking, predictability, and performance. Dominance is a measure of how in control a person feels and ranges from dominant to submissive [20]. Liking, as the name suggests, is a measure of how much a person likes the stimuli [23]. This is a useful metric since people can like negative valence emotions such as sadness and fear. A good example of this is the popularity of the horror movie genre. Fontaine et al. also proposed a fourth emotional dimension, predictability, as an important indicator of the surprise emotion along a spectrum [24]. This four-dimensional emotion model was the first to include the reasoning behind all six of the "basic" discrete emotions proposed by earlier researchers allowing for the blending of emotions within the four-dimensional space [24]. Table 1 below gives an example list of emotion models using continuous descriptors.

TABLE 1. CONTINUOUS EMOTION MODELS SUGGESTED BY RESEARCHERS

Year	**Paper**	**Continuous Emotion Categories**
1980	Russel [2]	Arousal, Valence (Circumplex model)
1988	Plutchik [21]	Emotion Wheel
1996	Mehrabian [20]	Arousal, Valence, Dominance
2005	Lee et al. [25]	Negative and Non-negative emotions
2006	Martin et al. [26]	Emotional Activation
2007	Fontaine et al. [24]	Arousal, Valence, Dominance, Predictability
2008	Vogt et al. [27]	Positive-Active, Negative-Active, Positive-Passive, Negative-Passive
2014	Hasan et al. [28]	Happy-Active, Happy-Inactive, Unhappy-Active, Unhappy-Inactive

2.1.2 Discrete Models

Discrete emotion models are the traditional way to describe emotions into separate defined categories such as anger, sadness, and happiness. Four authors in the 1990's produced seminal research into discrete emotion models: Ekman [29], Izard [30], Levenson [31], and Panksepp [18], and a good discussion of the rationale and analysis of the similarities and differences of these emotion models can be found in [32]. Table 2 below displays a list of different discrete emotion models developed and illustrates the variety of models available for use in applications such as affective computing. Question marks after an emotion label in the table refer to the researcher's uncertainty in the inclusion of that label in the emotion model.

TABLE 2. DISCRETE EMOTION MODELS SUGGESTED BY RESEARCHERS

Year	Paper	Discrete Emotion Categories
1992	Ekman [29]	Anger, Disgust, Fear, Happiness, Sadness, Contempt, Surprise
1992	Izard [30]	Anger, Disgust, Fear, Happiness, Sadness, Interest, Contempt?
1994	Levenson [31]	Anger, Disgust, Fear, Enjoyment, Sadness, Interest?, Love?, Relief?
1998	Panksepp [18]	Play, Panic/Grief, Fear, Rage, Seeking, Lust, Care
2005	Alm et al. [33]	Anger, Disgust, Fear, Happiness, Sadness, Positively Surprised, Negatively Surprised
2008	Strapparava et al. [34]	Anger, Disgust, Fear, Joy, Sadness, Surprise
2008	Gill et al. [35]	Anger, Disgust, Fear, Joy, Sadness, Surprise, Anticipation, Acceptance
2011	Balahur et al. [36]	Anger, Disgust, Fear, Joy, Sadness, Shame, Guilt
2012	Balabantaray et al. [37]	Anger, Disgust, Fear, Happiness, Sadness, Surprise
2012	Roberts et al. [38]	Anger, Disgust, Fear, Joy, Sadness, Surprise, Love
2012	Agrawal et al. [39]	Anger, Disgust, Fear, Happiness, Sadness, Surprise
2013	Sykora et al. [40]	Anger, Disgust, Fear, Happiness, Sadness, Shame, Surprise, Confusion
2013	Wang et al. [41]	Anger, Disgust, Fear, Joy, Sadness, Shame, Guilt
2013	Suttles et al. [42]	Anger, Disgust, Fear, Happiness, Sadness, Surprise, Trust, Anticipation
2013	Calvo et al. [43]	Anger, Disgust, Fear, Joy, Sadness

As can be seen from Table 2, there is a wide range of discrete emotion models which are similar but still have notable differences. This lack of consensus on a standardized emotion model introduces difficulty in any potential application of these models in other fields. While some emotions are easily related to each other such as joy, happiness, and enjoyment, others are more difficult to relate to each other since they describe similar feelings but have slightly different meanings such as disgust, shame, and guilt. It is also notable that all emotion models listed in Table 2 contain these emotions: anger, disgust, fear, joy, and sadness. If the reader would like to explore further information and discussion about discrete emotion models, they can refer to the comprehensive survey in [19].

2.2 Affective Datasets

There are multiple publicly available datasets that contain data of subjects while various emotions are elicited in an experimental setting. Table 3 lists many of the openly available emotional datasets which can be used to train and test emotion detection algorithms. A subset of these datasets also contain data regarding the subject's arousal and valence from the stimuli as described in the Circumplex model including the DEAP [23], SEMAINE [44], RECOLA [45], DECAF [46], SEWA [47], and CASE [1] datasets. The arousal and valence of these datasets are self-reported by the subjects during or after the presentation of the stimuli using discrete scales or continuous reporting mechanisms such as a joystick in the CASE dataset.

TABLE 3. POPULAR OPENLY AVAILABLE AFFECTIVE DATASETS

Dataset	**Year, Author**	**# of Subjects**	**Stimuli**	**Data Type**	**Emotion Model**
DEAP	2011, Koelstra et al. [23]	32	Videos	Physiological: EEG, GSR, Resp, Skin Temp, ECG, BVP, zEMG, tEMG, and EOG	Continuous (discretely self-reported after video): Arousal (1-5) Valence (1-5) Dominance (1-5) Liking (1-3)
SEMAINE	2012, McKeown et al. [44]	150	Simulated Conversation	Facial Video and Voice Audio Recordings	Continuous (labeled after the fact by experts) Valence Activation Power Anticipation/Expectation Intensity Discrete Fear, Anger, Happiness, Sadness, Disgust, Contempt, Amusement Epistemic States Interaction Process Analysis Validity
MAHNOB-HCI	2012, Soleymani et al. [48]	27	Videos	Facial Video, Audio, Eye Gaze,	Continuous: Arousal Valence Dominance Predictability Discrete: Disgust Amusement Joy Fear Sadness Neutral

TABLE 3. CONTINUED

Dataset	Year, Author	# of Subjects	Stimuli	Data Type	Emotion Model
RECOLA	2013, Ringeval et al. [45]	46	Collaborative Task	Video, Audio, ECG, GSR	Continuous: Arousal Valence Agreement Dominance Engagement Performance Rapport
DREAMER	2018, Katsigiannis and Ramzan [49]	23	Audio-Visual	EEG, ECG	Continuous: Arousal Valence Dominance
CASE	2019, Sharma et al. [1]	30	Videos	Physiological: ECG, BVP, EMG (3x), GSR, Resp, and Skin Temp	Continuous (continuously self-reported with joystick): Arousal (0-9) Valence (0-9) Discrete: Relaxed Bored Neutral Amused Scared
SEWA	2021, Kossaifi et al. [47]	398	Watching and Discussing Ads	Facial Video and Audio	Continuous: Valence Arousal Liking Agreement Disliking
DECAF	2015, Abadi et al. [46]	30	Music Videos and Movie Clips	MEG, NIR Facial Videos, hEOG, ECG, tEMG	Continuous: Arousal Valence Dominance
AMIGOS	2021, Miranda-Correa et al. [50]	40	Videos	EEG, ECG, GSR	Continuous: Valence Arousal Control Familiarity Liking Discrete: Neutral Disgust Happiness Surprise Anger Fear Sadness

2.3 Regression Machine Learning Models

2.3.1 Linear Regression

As the name suggests, linear regression creates a model that linearly correlates the inputs to the outputs. This allows the prediction of a value along a continuous range given a specific input by creating a linear best fit line. There are a couple of different methods for creating this best fit line including simple linear regression using statistics values such as mean, standard deviation, correlations, and covariance, ordinary least squares which minimizes the residual sum of squares, gradient descent which optimizes the model's coefficients by iteratively minimizing the error of the model using the training data, and regularization which uses the ordinary least squares method but also attempts to reduce the complexity of the model by optimizing the coefficients through various methods [51]. Fig. 3 below is an example of linear fit lines for two sets of data.

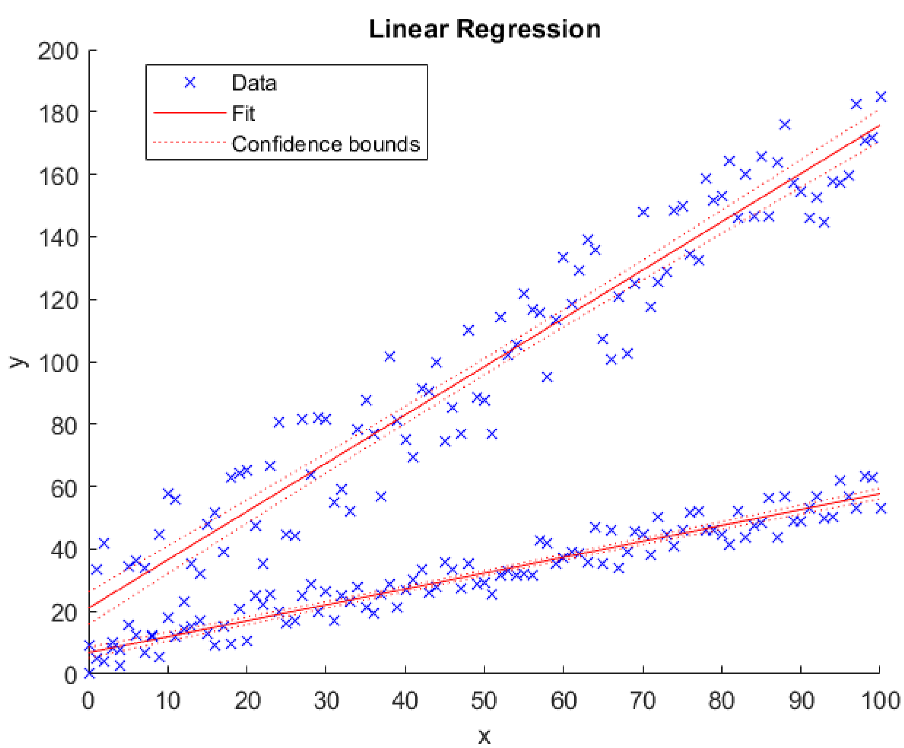


Fig. 3. Example linear regression best fit lines on two datasets.

2.3.2 Random Forest Regressor

The random forest regression is functionally constructed in the same way the random forest classifier models are constructed. The theory behind the random forest model structure is described in detail in the Classification Machine Learning Models section below. The major difference between a random forest classifier and regressor is that in the classifier, the decision is based on a majority vote of the decision trees for which class the input feature set is classified, and in the regressor, the average of decision tree outputs is calculated and taken as the value prediction from the regression [52]. Fig. 4 below gives an example of a single decision tree used for regression. The random forest regressor is an ensemble model using multiple of these decision tree regressors to determine the final regression output.

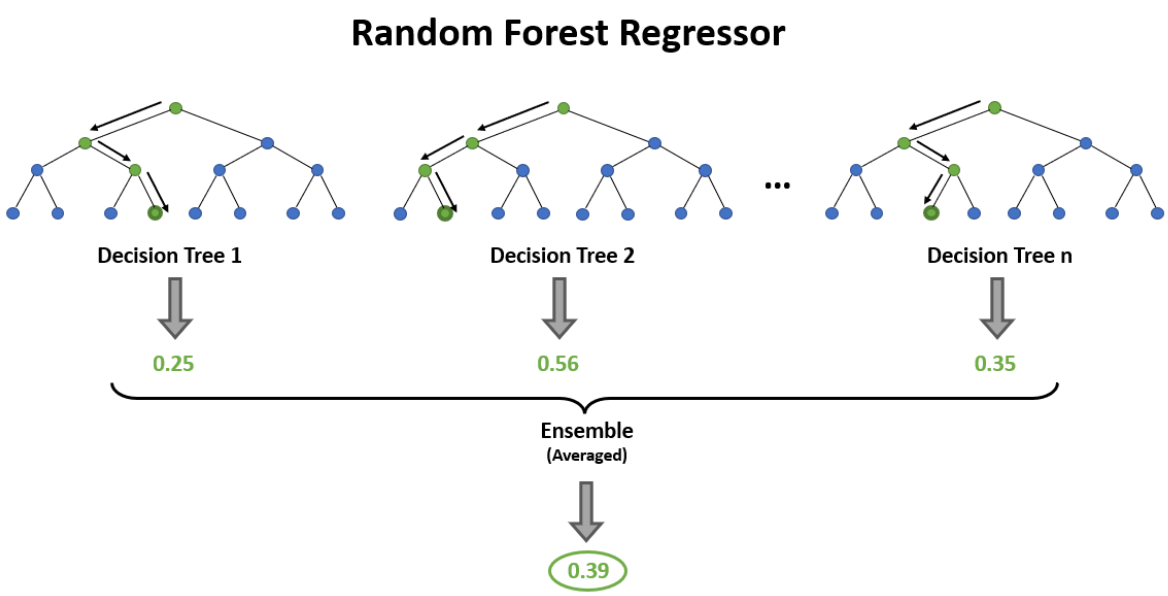


Fig. 4. Simplified Diagram of Random Forest Regressor.

2.3.3 Support Vector Regressor

Support vector regression uses the same principles as support vector machines used in classification problems by creating a hyperplane in higher dimensional spaces. A more detailed explanation of support vector machines in general is given in the Classification Machine Learning Models section below. However, instead of finding a hyperplane which attempts to maximize the separation of the datapoints between datapoint classes, the regressor attempts to find a hyperplane that contains, or touches, as many datapoints in the dependent variable as possible. This hyperplane is then used to predict new datapoints of the dependent variable. Fig. 5 below shows example hyperplanes produced by SVRs that contains a majority of the datapoints in the dataset using three different kernel functions.

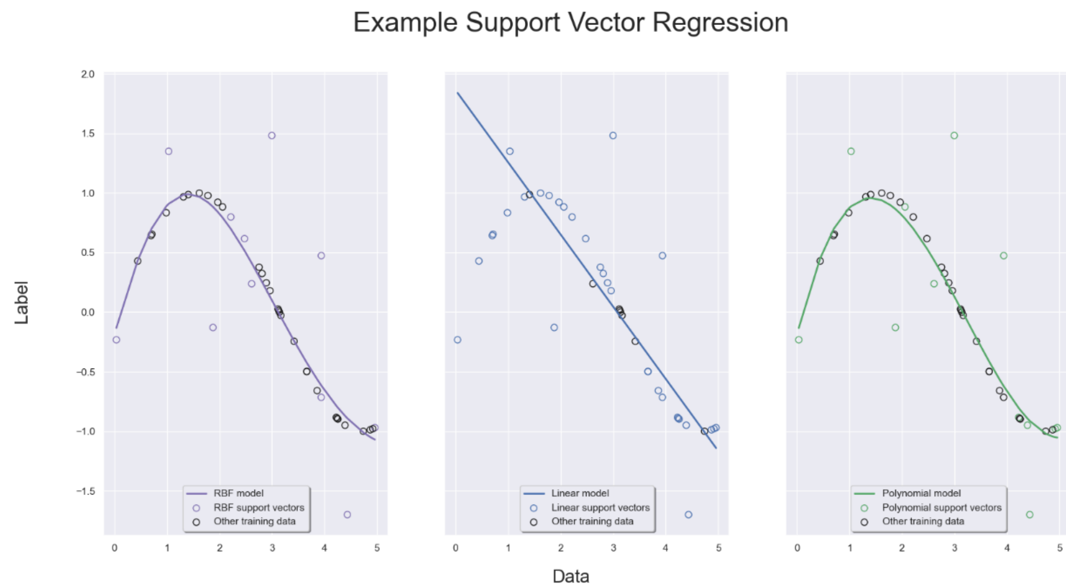


Fig. 5. Example hyperplanes created by different SVR kernels.

2.3.4 Adaboost Regressor

Adaboost, or "Adaptive Boosting", regression is a type of ensemble learning which uses the output of multiple weak learners to create a single strong learner [53]. The Adaboost algorithm in particular takes the output of multiple decision trees with a single split, known as "decision stumps", sequentially where each decision tree's output is used to improve the next decision tree's output. It is interesting to note that random forest models also use decision trees, but instead of running the decision trees in "parallel" to each other and aggregating the outputs of the decision trees at the end, the Adaboost model runs the decision trees sequentially and improves each tree in each iteration. Each decision is a separate regressor which is improved on itself for a predetermined amount of times set by the model designer. The resulting regressor is the ensemble regression model used. Fig. 6 illustrates the multiple decision stumps being used to improve the performance of the ensemble decision stump.

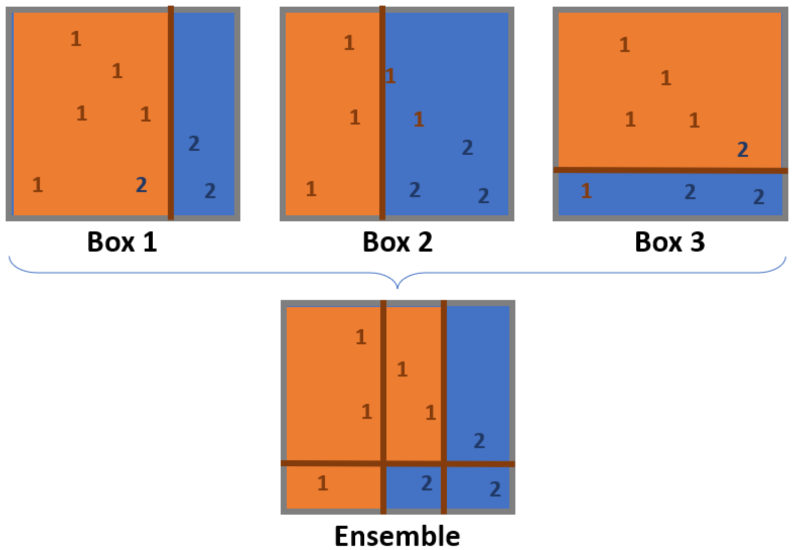


Fig. 6. Adaboost aggregation of multiple decision stumps (Box 1-3) into the final output (Box 4).

2.3.5 XG Boost Regressor

Extreme Gradient Boosting (XG Boost), like Adaboost, is a type of ensemble boosting model. The base regression model is a decision tree that is improved upon in each consecutive decision tree. Unlike Adaboost, XG Boost uses a loss function and gradient descent optimization to improve the performance of each decision tree [54]. The algorithm itself is designed to be scalable and fast which enables the use of the gradient boosting method on very large datasets.

2.4 Classification Machine Learning Models

2.4.1 Neural Network

Neural networks are composed of input, hidden, and output layers each with a set of nodes that are ‘connected’ with sets of weights and biases. Each node contains a nonlinear activation function such as the sigmoid or rectified linear unit (ReLU) functions which filter the output of that node non-linearly. If the non-linear activation functions were not present in the nodes, increasing the number of layers would have no effect on the accuracy of the neural network. Fig. 7 below shows the general structure of a neural network with an input layer, hidden layer(s), an output layer, nodes, weights, biases, and activation functions.

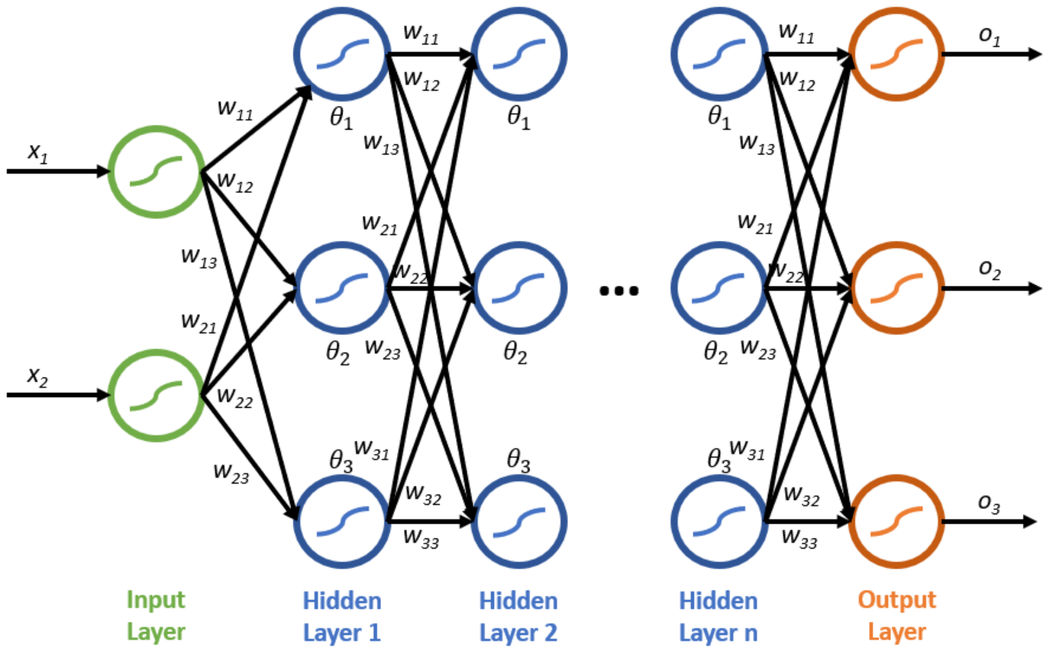


Fig. 7. Example Neural Network Structure.

Each circle represents a node where the first column is the input layer, the middle columns are the hidden layer(s), the last column is the output, x_i are the inputs (features), w_{ij} are the weights, θ_i are the biases, o_i, are the outputs (classification or regression), and the symbols inside the nodes represent the activation function.

2.4.2 Random Forest

The Random Forest model is a common, robust machine learning algorithm used for supervised learning of datasets in classification and regression scenarios. It is an ensemble method that combines the classification prediction from multiple decision tree classifiers based on a majority voting system. It also incorporates a random subsampling through the replacement of the original feature set for each decision tree to reduce the correlation between trees. This greatly reduces the variance between prediction outputs without increasing bias making random forest models robust to oversampling and accurate.

The functional component of the random forest model, the decision tree model, is a classification algorithm that makes a decision for each feature in a datapoint's feature vector based on a predefined function such as Gini Impurity or Information Gain. The structure of a decision tree model is shown in Fig. 8 below.

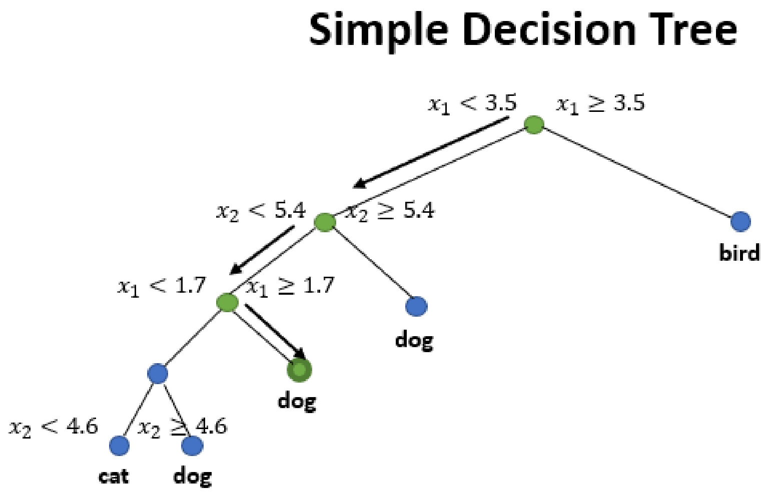


Fig. 8. Example decision tree showing node and branch structure for classifying data. The features used for classification and decision-making are seen in the branches labelled as x_n.

The Gini Impurity function is calculated as the sum of probabilities that a class, i, will be chosen, f_i, times the probability of misclassifying that class, $1 - f_i$. The equation for Gini impurity of a dataset with J classes is

$$I_G(f) = \sum_{i=1}^{J} f_i(1 - f_i) = \sum_{i=1}^{J}(f_i - f_i{}^2) = \sum_{i=1}^{J}(f_i - f_i{}^2) = \sum_{i=1}^{J} f_i - \sum_{i=1}^{J} f_i{}^2 = 1 - \sum_{i=1}^{J} f_i{}^2 = \sum_{i \neq k} f_i f_k \qquad (1)$$

The branch with the minimum impurity value is then chosen to continue the classification until a node is reached where the impurity equals zero, i.e. every case in the node is the same class. The decision tree then classifies that datapoint in that class.

The other possible decision function, information gain, aims to minimize the complexity of the decision tree by finding the split that minimizes the amount of information needed in the child node to make a decision on the class of the datapoint. To define this amount of information, the entropy of a node is calculated and defined by

$$H(T) = I_E(p_1, p_2, \cdots, p_n) = \sum_{i=1}^{J} p_i \log_2 p_i \qquad (2)$$

The information gain from parent to child node is then calculated as the difference between the parent node's entropy and the child node's entropy given by

$$IG(T,a) = H(T) - H(T|a) \tag{3}$$

The feature with the highest information gain, or difference between parent and child node entropy, is then chosen to split the node on.

The Random Forest algorithm then uses a method called feature bagging to aggregate the classifications of multiple decision trees to reduce the variance of predictions by statistically validating them using multiple models. Fig. 9 below shows this aggregation and the equation used to make the final classification prediction for the ensemble model.

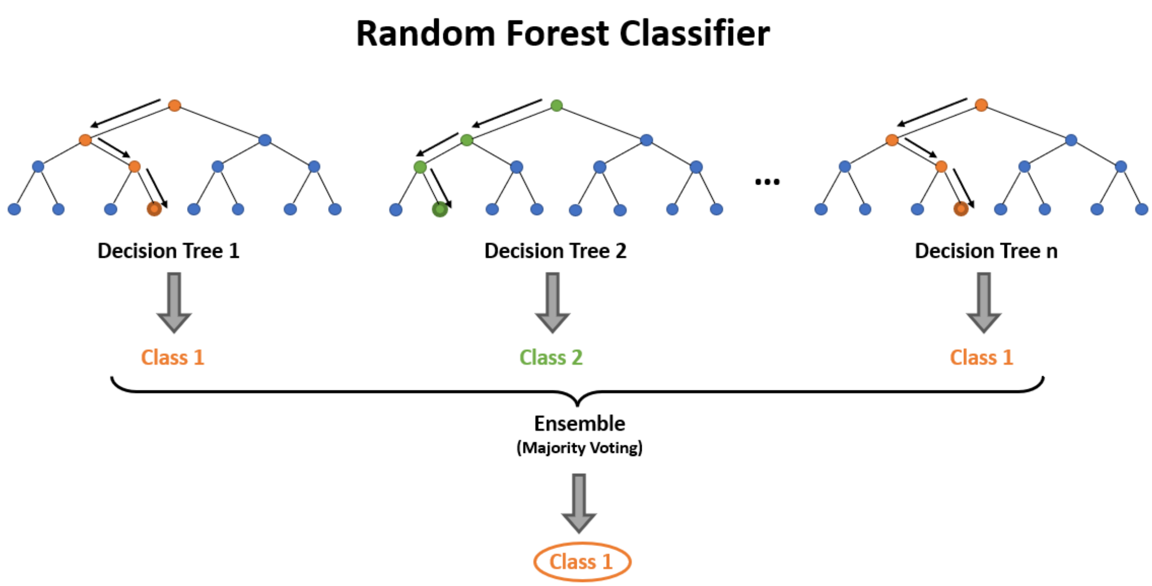


Fig. 9. Simplified Diagram of Random Forest Classifier.

For each decision tree in the random forest, a subset of features is chosen from the full feature set to classify each observation. This ensures that each tree is sufficiently different from the rest to prevent overfitting of the training data. It also decreases variance while keeping the bias relatively low. This makes random forests more robust to overfitting and accurate than a single decision tree model.

2.4.3 Support Vector Machines (SVM)

SVMs are a common classification algorithm that classifies data by creating hyperplane boundaries in multidimensional space. The optimal hyperplane is defined as the plane with the farthest distance between the datapoints of two separate classes. For multiclass classification, a set of hyperplanes is created to separate the distributions of multiple sets of datapoints.

A kernel function, $k(x, y)$, is used to project the data into a higher-dimensional space for easier separation by a hyperplane. Fig. 10 below displays a projection of a dataset into a higher dimensional space through a kernel function.

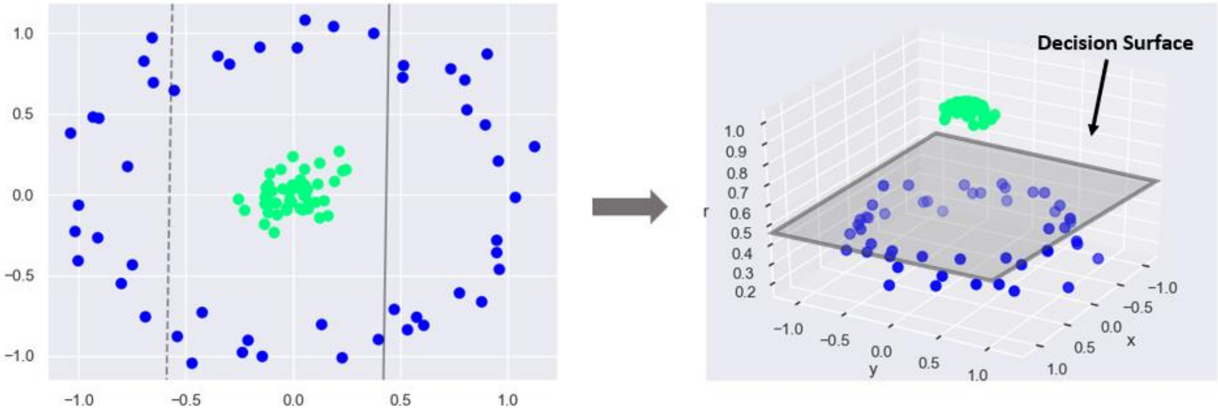


Fig. 10. Projection of dataset into higher dimensional space by a radial basis kernel function, r.

This kernel function can be linear or non-linear depending on the dataset and application. A nonlinear kernel function is more accurate than linear kernels since the boundary is more flexible in its distinguishing between two datasets, but it is much more computationally expensive. A linear kernel is less accurate and much less computationally expensive, but with enough data, the accuracies of a linear kernel are similar to those of a non-linear kernel. For this reason, linear-kernel SVMs are useful for large volumes of data, and non-linear kernel SVMs are more applicable to smaller datasets.

The hyperplane is defined by the location of the datapoints of one class closest to the datapoints of another class. These datapoints are denoted as *support vectors*. As stated in [55], by convention, the formal equation of a hyperplane is defined as

$$|\beta_0 + \beta_X^T| = 1 \tag{4}$$

where β_0 is known as the bias, β is the weight vector, and x is the vector/array of datapoints. Using (4) as the definition of a hyperplane, we can compute the distance from the hyperplane to a support vector as

$$\text{distance}_{\text{support vectors}} = \frac{|\beta_0 + \beta_X^T|}{||\beta||} = \frac{1}{||\beta||} \tag{5}$$

The margin, or distance between the two closest datapoints in different classes, is calculated as

$$M = \frac{2}{||\beta||} \tag{6}$$

This margin needs to be maximized to obtain the optimal hyperplane to separate the data. The following equation maximizes the margin by minimizing a function, $L(\beta)$, such as

$$\min_{\beta,\beta_0} L(\beta) = \frac{1}{2}||\beta||^2 \text{ subject to } y_i(\beta_{x_i}^T + \beta_0) \geq \forall i \tag{7}$$

subject to the equation of the hyperplane with respect to the class, y_i, being greater than or equal to 1. This equation can be solved using Lagrangian optimization to find the optimal hyperplane to classify the data.

2.4.4 Convolutional Neural Network (CNN)

CNNs are a form of neural networks which implement a layer of convolution with a set amount of convolutional filters which detect features in an n-dimensional 'image' matrix. The filter size and number of filters can be chosen for each convolutional layer. The model calculates and updates weights and biases similar to a traditional neural network in each convolutional layer. The convolutional layer is followed by a nonlinear activation function such as the sigmoid or ReLU function.

There are also pooling, dropout, and fully connected layers. Pooling layers downsample the layer input by performing an operation such as max or average along an n-dimensional moving window. This preserves most of the information within the moving window while downsampling the data and increasing the speed of the algorithm. The dropout layer is a method of preventing overfitting by randomly dropping out nodes in the network [56]. This prevents the nodes from correlating with each other too much during training and overfitting the data [56]. The fully connected layer reduces the dataset to a 1-dimensional vector corresponding to the classification of the images. The final layer of a CNN is a vector with the number of classes where each node corresponds to the probability that that datapoint belongs to that class. The max node in that layer is the class that is assigned to that datapoint. Fig. 11 below visualizes an example layout of a CNN.

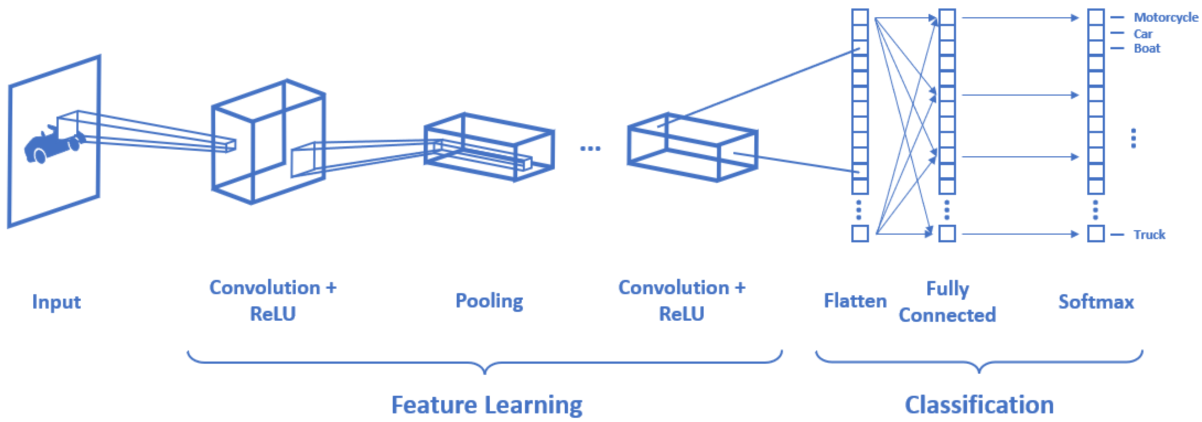


Fig. 11. Example architecture of a CNN showing the different types of layers.

2.4.5 Long Short-Term Memory (LSTM)

Long Short-Term Memory (LSTM) models are a type of Recurrent Neural Network (RNN) model which are special types of Neural Networks that can gain insight into data at the current timestamp using data from previous timesteps as contextual information [57]. This allows RNNs to use temporal information to improve performance for applications such as speech detection, music composition, handwriting detection, grammar insights, and other time-dependent datasets. LSTMs are a special type of RNN that helps overcome the technical issues with simple RNNs of vanishing gradients and exploding gradients [56]. The vanishing gradient problem is an issue with RNNs where the influence of an input either decays or blows up exponentially as it cycles around the network's recurrent layers, and the LSTM is designed specifically to alleviate this issue [56]. This allows LSTMs to perform better with longer time lags between the current datapoint and previous datapoints to gain insight on data further back in time.

A general LSTM node consists of an input gate, output gate, and forget gate [58]. Fig. 12 below illustrates a single LSTM cell within a neural network structure:

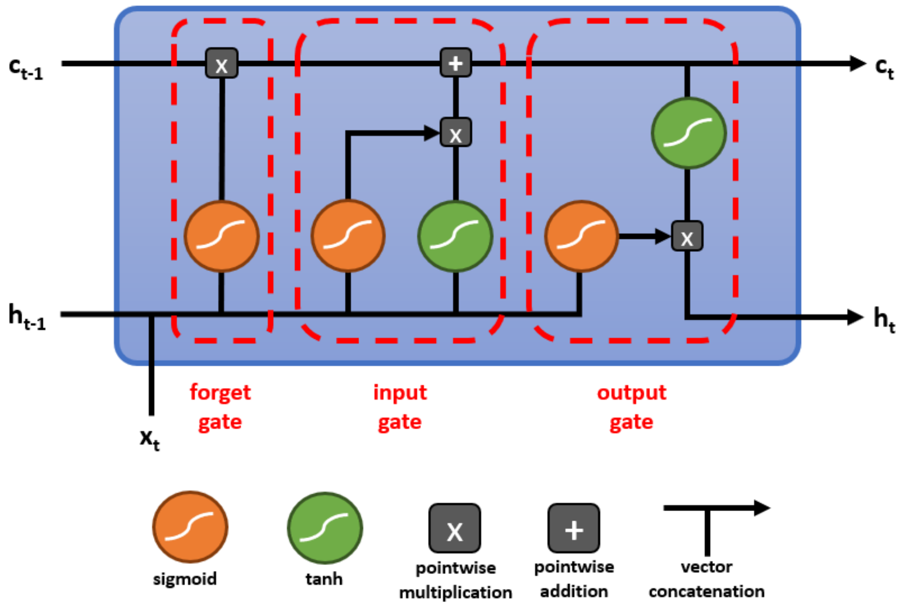


Fig. 12. Single LSTM node layout illustrating mathematical operations performed on the input to achieve output.

The terms c_{t-1} and c_t represent the previous and current cell state, h_{t-1} and h_t represent the previous and current hidden state, and x_t represents the new input value(s) for this timestep. The first gate, forget gate, decides whether the information should be kept or forgotten through a sigmoid function. The next gate, input gate, uses a tanh function to regulate the values between -1 and 1 as well as a sigmoid function to determine the importance of the output of the tanh function. The outputs of the forget and input gates are added together to produce the next cell state. The third gate, output gate, decides what the next hidden state should be by taking the tanh of the new cell state and multiplying it by the same output of the sigmoid function used in the previous gates. This product is then the new hidden state for this timepoint.

2.5 Literature Review

The field of affective computing is extensive and diverse due to the increasing availability of quality datasets, wearable sensor technology, and processing capability. Affective computing has been investigated since the late 1980s, and some of the first research into affective computing came from the University of Iowa and Ohio State University using facial electromyography to differentiate valence and affective state [59]. Dr. Picard’s Media Laboratory at the Massachusetts Institute of Technology (MIT) was also responsible for many of the first seminal studies published on affective computing [11, 60-62]. It is difficult to compare results from affective computing and emotion detection research studies due to the varying nature of the emotional stimuli, methodologies, and emotion labeling/model used. However, emotion detection research can be broadly categorized into similar methodologies such that a comparison of the classification accuracies or regression errors can be made. This literature review categorizes the extensive research of emotion detection into three groupings: multi-class discrete emotion label classification models, multi-class arousal/valence classification models, and arousal/valence regression models. It would not be useful or prudent to review all research performed on affective state detection to date, so this review describes a sampling of papers in order to provide insight into the evolution of affective state detection research from the late 1980s to the present.

2.5.1 Discrete Emotion Classification

This emotion detection method uses classification models to classify data collected from human subjects into discrete emotion labels. The type of data, features extracted, models, number of emotions, and descriptors of emotions all vary across subjects. Table 4 below describes a sampling of research papers that shows the evolution of emotion classification models from some of the first models developed in the late 1990s up to the present. In the late 1990s, the Media Lab at MIT published some of the first seminal research papers on discrete emotion detection. These papers were published before the rise in popularity of machine learning techniques and used statistical techniques such as the Maximum a Posteriori (MAP) model to classify 8 emotions with a 48.8% accuracy. While this result is low by today's standards, they proved that there was in fact a correlation between physical physiological signals and the abstract theories of emotion which laid the groundwork for future affective computing research. Over the next two decades, emotion classification research evolved to use popular machine learning algorithms with some accuracies achieving up to 97% with a three-class model using an SVM classifier in 2020 [63]. However, this classification was trained and tested on a single subject's data making the model subject dependent, and an important distinction for comparing accuracies across different papers is the creation of subject dependent vs. independent models. A model created and tested on the same subject's data is significantly easier to create than a model created on a set of subjects' data and trained on a completely different set of subjects' data. Most research performed to date is subject-dependent, meaning the models have little applicability outside of use by a single subject.

TABLE 4. DISCRETE EMOTION CLASSIFICATION RESEARCH REVIEW

Year	Author	Data Set	# of Feat.	Feature Selection	Classification Model	Emotions Classified	Results
1998	Vyzas and Picard [62]	EMG, BVP, GSR, Resp	6	None	MAP	Neutral (N), Anger (A), Hate (H), Grief (G), Platonic Love (P), Romantic Love (L), Joy (J), Reverence (R)	8-class: 48.8%, 5-class (NAGJR): 71.0%, 4-class (NAGR): 72.5%, 4-class (AGJR): 72.5%, 3-class (AGR): 83.3%, 3-class (AJR): 88.3%
2000	Healey [11]	EMG, BVP, ECG, Resp, EMG	11	8-emotion: Fisher Projection and Sequential Floating Feature Selection (SFFS)	8-emotion: K-Nearest Neighbors (KNN) 3-emotion: Linear and Quadratic	No Emotion, Anger, Hate, Grief, Platonic Love, Romantic Love, Joy, Reverence	8-class: 81.3% 3-class: 87.0%
2005	Wagner et al. [64]	EMG, ECG, GSR, Resp	32	SFS	Linear Discriminant Function (LDF)	Joy, Anger, Sadness, Pleasure	4-class: 92.1%
2005	Herbelin et al. [65]	EMG, BVP, GSR, Resp, Skin Temp, Arousal/ Valence labelled by subject	36	Fisher LDA (reduced to 2 features)	KNN	Neutral, Fear, Boredom, Joy, Exaltation	5-class: 24.0%
2008	Maaoui and Pruski [66]	EMG, BVP, GSR, Resp, Skin Temp	6	None	SVM w/ linear kernel	Amusement, Contentment, Disgust, Fear, Neutral, Sadness	6-class: 88.0%
2014	Verma and Tiwary [67]	EEG, EMG, EOG, BVP, GSR, Resp, Skin Temp	25	None	SVM, MLP, KNN, MMC	Terrible, Love, Hate, Sentimental, Lovely, Happy, Fun, Shock, Cheerful, Depressing, Exciting, Melancholy, Mellow	Terrible: 80.9% Love: 82.2% Hate: 79.8% Sentimental: 82.5% Lovely: 81.6% Happy: 82.8% Fun: 79.8% Shock: 79.7% Cheerful: 80.3% Depressing: 85.5% Exciting: 83.6% Melancholy: 77.7% Mellow: 82.5%

TABLE 4. CONTINUED

Year	Author	Data Set	# of Feat.	Feature Selection	Classification Model	Emotions Classified	Results
2014	Wen [68]	BVP, GSR	3	None	Random Forest	Amusement, Anger Grief Fear Neutral	5-class: 74.0%
2019	Albrai-kan et al. [69]	BVP, GSR, Skin Temp, and MAHNOB	Raw Data	None	Weighted Multi-Dimensional Dynamic Time Warping (WMD-DTW), KNN	Neutral, Cheer, Sadness, Erotic, Horror	5-class: 65.6%
2019	Bălan et al. [70]	EEG, EOG, EMG, BVP, GSR, Resp, Skin Temp	Raw Data	Fisher, PCA, SFS	DNN, SVM, RF, LDA, KNN	Binary (yes/no): Anger, Joy, Surprise, Disgust, Fear, Sadness	Anger: 98.3% Joy: 100% Surprise: 96.0% Disgust: 95.0% Fear: 90.8% Sadness: 90.8%
2020	Domín-guez-Jiménez [63]	BVP, GSR	27	Random Forest Recursive Feature Elimination	SVM	Amusement, Sadness, Neutral	3-class: 97.0%
2020	Liu et al. [71]	GSR	6	None	3-Layer NN	Anger, Disgust, Fear, Happy, Surprise, Sad, Neutral	7-class: 42.1%
2021	Oh et al. [72]	GSR and Predicted Arousal/Valence from Facial Images	Raw Data	None	DNN named “Sensor Fusion Emotion Recognition (SFER)”	Neutral, Happy, Excited, Fearful, Agony, Depressed, Bored, Relieved	8-class: 89.0%

2.5.2 Arousal/Valence Classification

Another popular type of emotion classification research methodology is classifying arousal and valence in the Circumplex model of affect into different discrete classes along the continuous range of arousal and valence. Most papers used two classes for arousal: low and high arousal, and two classes for valence: negative and positive valence. Some papers split the continuous ranges of arousal and valence into three classes: low/middle/high arousal and negative/neutral/positive valence. Most papers used two separate classification models: one for arousal and another for valence. Some of them, however, used a single model to classify four classes: low arousal (LA), high arousal (HA), negative valence (NV), and positive valence (PV). As can be seen, there are multiple classification methodologies of varying practicality and training difficulty. The models in this category are similar enough to be adequately compared since the resulting classes are the same or very similar. One of the first emotion recognition papers released used this methodology to determine a correlation between physiological signals and emotions in the 1980s [73]. Since then, hundreds of research papers have been published for classifying arousal and valence into discrete classes. The accuracies of these papers range from the 50s percentage in the early 2000s to the mid-90s percentage in 2019 in a paper by Albraikan et al. [69]. Table 5 below gives a sampling of papers using this arousal/valence classification methodology and shows the progression of modeling techniques and accuracies.

TABLE 5. AROUSAL/VALENCE CLASSIFICATION RESEARCH REVIEW

Year	Author	Data Set	# of Feat.	Feature Selection	Classification Model	Emotions Classified	Results
1986	Cacioppo et al. [59]	Facial EMG	6	None	Multivariate Analysis	Pos./Neg. Valence	Correlation was Statisticall y Significant
1998	Healey and Picard [61]	EMG, BVP, GSR, Resp	11	None	Fisher Linear Discriminate Projection	Low/High Arousal Pos./Neg. Valence	Low Arousal: 80.0% High Arousal: 88.0% Neg. Valence: 50.0% Pos. Valence: 82.0%
2000	Healey [11]	EMG, BVP, ECG, Resp, EMG	11	None	Linear and Quadratic	Low/High Arousal Pos./Neg. Valence	Low/High Arousal: 84.0% Neg,/Pos. Valence: 63.0%
2005	Wagner et al. [64]	EMG, ECG, GSR, Resp	32	SFS	LDF and NN	Low/High Arousal Pos./Neg. Valence	Low/High Arousal: 96.6% Neg,/Pos. Valence: 88.6%

TABLE 5. CONTINUED

Year	Author	Data Set	# of Feat.	Feature Selection	Classification Model	Emotions Classified	Results
2005	Herbelin et al. [65]	EMG, BVP, GSR, Resp, Skin Temp, Arousal/ Valence labeled by subject	36	Fisher LDA (reduced to 2 features)	KNN	Arousal Low/Middle/ High Valence Neg./Neutral/ Pos.	3-class Arousal: N/A 3-class Valence: 45.0%
2007	Jones and Troen [74]	BVP, GSR, Resp	11	None	NN	Scale from Low (1) to High (5) Arousal Scale from Pos. (1) to Neg. (5) Valence	Arousal: 67.0% Valence: 62.0%
2008	Khalili and Moradi [75]	EEG, BVP, GSR, Resp, Skin Temp	384	Genetic Algorithm	LDA and KNN	Valence: Calm (C), Positively Excited (PE), Negatively Excited (NE)	3-class (C vs PE vs NE): 51% (KNN) 2-class (PE vs NE): 70% (LDA and KNN)
2008	Gu et al. [76]	EMG, ECG, BVP, GSR	36	Genetic Algorithm	KNN, Fuzzy KNN, LDF, and Quadratic Discriminate Function (QDA)	Low/High Arousal Pos./Neg. Valence	Arousal: 77.0% Valence: 75.0%
2012	Koelstra et al. [23]	DEAP (original paper)	322	None	Decision Fusion	Low/High Arousal Pos./Neg. Valence	Arousal: 57.0% Valence: 62.7%
2013	Nogueira et al. [77]	Self-report: Arousal, Valence Physiolog-ical: GSR, Facial EMG, BVP	4	None	Decision Trees	Low/High Arousal Pos./Neg. Valence	Arousal: 98.2% Valence: 86.3%

TABLE 5. CONTINUED

Year	Author	Data Set	# of Feat.	Feature Selection	Classification Model	Emotions Classified	Results
2014	Torres-Valencia et al. [78]	DEAP	Raw Data	None	HMM	Low/High Arousal Pos./Neg. Valence	Arousal: 55% ± 3.9% Valence: 58% ± 3.9%
2017	Wiem and Lachiri [79]	MAHNOB-HCI	169	None	SVM	Low/High Arousal Pos./Neg. Valence	Arousal: 64.2% Valence: 65.0%
2017	Wiem and Lachiri [80]	MAHNOB-HCI	2	None	SVM w/ Gaussian Kernel	Calm/ Medium/ Activated Arousal, Unpleasant/Neutral/ Pleasant Valence	Arousal: 54.7% Valence: 57.4%
2017	Kawde and Verma [81]	DEAP	Raw Data	None	DNN	High/Low Arousal Neg./Pos. Valence High/Low Dominance	Arousal: 70.7% Valence: 75.8% Dominance: 69.1%
2017	Henia and Lachiri [82]	MAHNOB-HCI	169	None	SVM	Calm/ Medium/ Activated Arousal, Unpleasant/Neutral/ Pleasant Valence	Arousal: 59.6% Valence: 57.4%
2018	Choi and Kim [83]	DEAP	Raw Data	None	LSTM	High/Low Arousal Neg./Pos. Valence	Arousal: 74.7% Valence: 78%
2018	Sarabadani et al. [84]	ECG, GSR, Resp, Skin Temp	23	None	Ensemble of: KNN (k=3), LDA, SVM (linear), SVM (poly) SVM (RBF)	HA/NV vs. HA/PV LA/NV vs. LA/PV	HA/NV vs. HA/PV: 78.1 ± 11.7% LA/NV vs. LA/PV: 84.5 ± 9.8%

TABLE 5. CONTINUED

Year	Author	Data Set	# of Feat.	Feature Selection	Classification Model	Emotions Classified	Results
2018	Ali et al. [85]	MAHNOB	25	None	Cellular Neural Network	LA, HA, NV, PV	4-class: 89.4%
2018	Ayata et al. [86]	DEAP	22	mRMR	Random Forest, SVM, Logistic Regression	High/Low Arousal Neg./Pos. Valence	Arousal: 73.1% Valence: 72.2%
2019	Albraikan et al. [87]	BVP, GSR, Skin Temp, and MAHNOB	Raw Data	None	Weighted Multi-Dimensional Dynamic Time Warping (WMD-DTW), KNN	Arousal Calm/Medium/Activated Valence Unpleasant/ Neutral/ Pleasant	3-class Arousal: 94% 3-class Valence: 93.6%
2020	Liu et al. [71]	GSR	6	None	3-Layer Neural Network	Anger, Disgust, Fear, Happy, Surprise, Sad, Neutral	High/Low Arousal: 68.7% Pos./Neg. Valence: 72.7%
2020	Li et al. [88]	AMIGOS	LSTM -RNN	None	DNN	High/Low Arousal Neg./Pos. Valence	Arousal: 82.5% Valence: 77.8%
2020	Baghizadeh et al. [89]	MAHNOB-HCI	Poin-caré Map	None	KNN, SVM, MLP	High/Low Arousal Neg./Pos. Valence	Arousal: 82.2 ± 4.7% Valence: 78.1 ± 3.4%

2.5.3 Arousal/Valence Regression

The last category of emotion detection methodologies discussed in this literature review contains the least amount of research papers published and corresponds to using regression models to predict arousal and valence along continuous spectrums. Since regression models predict values continuously, accuracy cannot be used to determine the performance of the models. Root mean squared error (RMSE), mean squared error (MSE), and mean absolute error (MAE) are all metrics that give insight into the effectiveness of a regression model. In general, the lower the error value, the better the model predicts the dependent variables since the aggregated error between the predicted and true values is lower. In Table 6 below, the RMSE, MSE, and MAE metrics are given for research that explored regression for arousal and valence.

A few studies in Table 6 detected emotions from data other than physiological signals, but they were included due to the similarity in methodology to the one presented in this work. Han et al. detected emotion in pop songs by mapping the discrete emotions labeled by the All Music Guide [90] to the arousal and valence two-dimensional space themselves [91]. This mapping was then used as ground truth arousal and valence values for training a support vector regressor (SVR) on features extracted from the songs to predict arousal and valence in cartesian or polar coordinates [91]. This predicted arousal and valence were then remapped back into discrete emotions and the accuracy of the original emotion labels versus the predicted emotions was 94.55% using SVR and polar arousal and valence coordinates [91]. Another study that used a similar methodology as the one presented in this work was by Nogueira et al. in classifying high/low arousal and negative/positive valence using predicted arousal and valence from regression models [77]. The regression results are presented in Table 6 below and the final classification accuracies are presented in Table 5 above. The most recent study in this review, written in 2021 by Oh et al., uses a similar method to the one explored in this work where they use a regression model to predict arousal and valence from facial expressions which are then used as input into a classification model for classifying discrete emotions [72]. The results of the final classification model are noted in Table 5 above.

It is difficult to compare arousal and valence regression studies due to their use of different performance metrics such as RMSE, MSE, MAE, and R^2. The only metric used by more than one study in the review in Table 6 below was MAE, and the best MAE for arousal was 1.49 ± 0.42 and for valence was 1.56 ± 0.36 both by Soleymani et al. [92].

TABLE 6. AROUSAL/VALENCE REGRESSION RESEARCH REVIEW

Year	Author	Data Set	# of Feat.	Feature Selection	Regression Model	Emotions Classified	Results
2009	Han et al. [91]	Pop songs labeled with emotions (All Music Guide [90])	7	None	Support Vector Regression (SVR)	Distance from origin and Angle (polar coordinates) in Arousal/Vale-nce space	Distance: MSE: 0.025 Angle: MSE: 0.098

TABLE 6. CONTINUED

Year	Author	Data Set	# of Feat.	Feature Selection	Classification Model	Emotions Classified	Results
2011	Soleymani et al. [92]	Self-report: Arousal, Valence, Dominance, Liking Physiological: EEG, EMG, BVP, GSR, Resp, Skin Temp	177	None	Linear Ridge Regressor	Arousal 0-9, Valence 0-9, Liking 0-9, Dominance 0-9	Arousal: MAE: 1.49 (0.42), Valence: MAE: 1.56 (0.36), Liking: MAE: 1.51 (0.49), Dominance: MAE: 1.66 (0.46)
2013	Nogueira et al. [77]	Self-report: Arousal, Valence Physiological: GSR, Facial EMG, BVP	4	None	SC-Arousal: linear HR-Arousal: 3rd-degree polynomial zEMG-Valence (positive): 3rd-degree polynomial cEMG-Valence (negative): 3rd-degree polynomial HR-Valence: 3rd-degree polynomial	SC-Arousal HR-Arousal zEMG-Valence (positive) cEMG-Valence (negative) HR-Valence	SC-Arousal R^2: 0.90 ± 0.038 HR-Arousal R^2: 0.74 ± 0.089 zEMG-Valence (positive) R^2: 0.92 ± 0.016 cEMG-Valence (negative) R^2: 0.95 ± 0.075 HR-Valence R^2: 0.96 ± 0.064
2014	Torres-Valencia et al. [93]	DEAP	16	Recursive Feature Eliminat-ion (RFE)	Multiple Output Support Vector Regression (M-SVR)	Arousal Valence	Arousal: RMSE: 0.240 ± 0.024 MAE: 0.203 ± 0.020 Valence: RMSE: 0.252 ± 0.026 MAE: 0.213 ± 0.021
2021	Oh et al. [72]	Facial Expression	Raw Data (Feat. from DNN)	None	SE-ResNeXt	Arousal Valence	Arousal: RMSE: 0.408 Valence: RMSE: 0.373

2.6 Limitations of Current Studies

An important distinction when comparing emotion detection models is whether the model was trained and tested on data from multiple subjects rather than training and testing on data from a single subject. Most research performed early on in emotion detection only used a single subject's data to create a model. This produces models which are not generalizable for use with other people, and thus the practicality of the models is greatly reduced. Even if a model is trained and tested on multiple subjects' data, it is important to separate the subjects whose data is in the training set from the subjects whose data is in the testing set. This gives a more accurate indication of the generalizability of the model by determining the testing accuracy of the model on subjects whose data was never seen before by the model.

Current emotion detection research has reached accuracies up to 97% for discrete emotion detection (3-class) by Dominguez-Jimenez et al. [63], 98.2% for discrete arousal (2-class) by Nogueira et al. [77], 93.6% for valence classification (3-class) by Albraikan et al. [87], and 1.49 ± 0.42 MAE for arousal and 1.56 ± 0.36 MAE for valence regression by Soleymani et al. [92] (although it is difficult to compare regression results due to the different metrics used among studies). The 97% 3-class accuracy in Dominguez-Jimenez et al. is impressive, but the models generated were *subject-dependent* meaning the models only provide that accuracy when used on the single subject it was trained on. The arousal and valence classification accuracies are also high with mid to high 90s percentages, but the practicality of using an arousal and valence classification is limited in applications of direct emotion feedback to a user due to the lesser-known emotion definitions of arousal and valence among the public. It can, however, be used in emotional feedback applications where a system is intending to improve a user's affective state by detecting, for example, negative valence and responding to help the user feel more positive. In Nogueira et al., the end classification model is subject-independent, but the regression model used to generate the predicted arousal and valence features is subject-dependent which necessitates a "calibration procedure" where a new user would need to self-report arousal and valence for a period to retrain a regression model which would be specific to them [77]. The models produced by Albraikan et al. are subject-independent with relatively high accuracies for arousal and valence classification, but the utility of arousal and valence classification in real-world applications is limited. Arousal and valence regression have limitations similar to arousal and valence classification models, but they can also apply to emotional feedback applications.

Most models use physiological sensors such as EEG, MEG, EMG, and NIRs which are difficult to implement in everyday situations outside of controlled lab settings with today's technology. Another limitation in the affective research field, in general, is the lack of a standardized approach for describing emotions. This makes it difficult to compare results across different methodologies. The lack of publicly available, well-acquired emotion data is also a limitation, but this has improved recently in the past couple of years with the addition of a decent number of openly available datasets as described in Table 3 above.

These datasets, however, use different emotion description models from one another which reduces their practicality in terms of using data from multiple datasets in the same model. For example, if one dataset uses bored, relaxed, neutral, amused, and scared emotion labels, and another uses amusement, contentment, disgust, fear, neutral, and sadness emotion labels, it is up to the classification model designer to match emotion labels with each other from these two datasets if they would like to use data from both in their model training and testing. While these emotion labels are similar, they are not the same, and this introduces difficulty in using more than one dataset for model generation which effectively limits the amount of data available for model training even as more datasets are released.

Another limitation in current studies involving self-reported emotion labeling is the inherent bias in the self-reporting of these values by the subjects. An example of this is a subject’s knowledge that a stimulus is “supposed” to elicit a certain emotion or reaction, so the subject is more inclined to self-report that response. This is a psychological phenomenon known as confirmation bias [94]. Another example is the desire to display socially acceptable behavior and responses such as when asked about emotions like erotica and charity [95]. The quality of the model is reflected in the quality and truthfulness of the self-reporting by the subjects [87]. This is a difficult limitation to overcome since researcher-defined arousal and valence “ground truth” values are also limited by the subjective nature of emotion elicitation among subjects. There exists no “perfect” method for defining ground-truth emotion values, but improvements are being made to produce more accurate arousal and valence labeling techniques such as the method used in the CASE dataset with the *Joystick-based Emotion Reporting Interface* (JERI) annotation device whose data is utilized in this work [1, 96].

CHAPTER 3

MATERIALS AND METHODS

Fig. 13 gives an overview of the emotion detection model generation process described in this work. Time-synced physiological data, continuously self-reported arousal and valence values, and emotion labels from the publicly available CASE dataset are used as inputs and target labels for generating the ensemble model. The physiological data is windowed into 10-second-long segments with a 1-second window stride, low-pass/band-pass filtered, and feature extracted to produce a feature set of physiological features. The arousal and valence labels are resampled to produce a single average value for arousal and valence for a single window, reordered into low-to-high arousal and negative-to-positive valence respectively, and transformed into a Gaussian distribution with the same mean and standard deviation as the predicted arousal and valence labels to separate the labels more distinctly. The physiological features and "true" arousal and valence labels were then used to train regression models for predicting arousal and valence respectively (one regression model for arousal and one regression model for valence). The predicted arousal and valence features are then concatenated with the physiological feature set to create the combined feature set used to train and test the classification model which classifies each feature vector in the feature set as one of five emotions: amused, bored, neutral, relaxed, and scared.

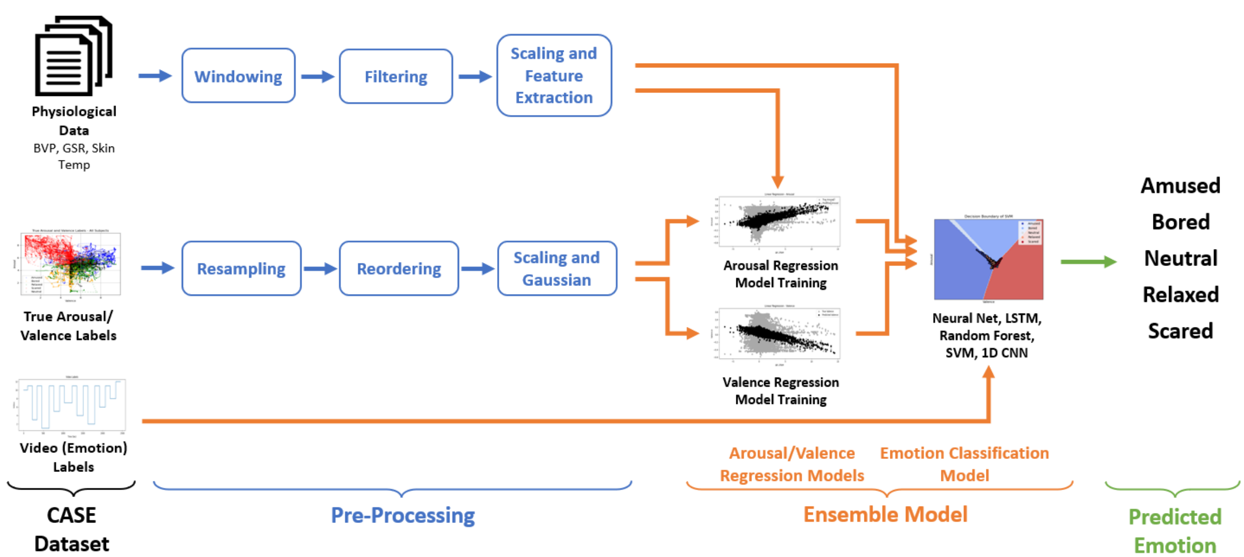


Fig. 13. Emotion Detection Ensemble Model Generation Methodology.

3.1 Dataset

The dataset chosen to train the emotion detection ensemble models, the *Continuously Annotated Signals of Emotion* (CASE) dataset, provides a uniquely labeled physiological dataset of elicited emotions since the device used to measure the subject's arousal and valence continuously was a two-dimensional joystick with the arousal and valence coupled together in the two dimensions called *Joystick-based Emotion Reporting Interface* (JERI) [96]. This allowed the subjects to simultaneously report arousal and valence continuously throughout the emotion elicitation while physiological signals were being recorded. Fig. 14 below is from Sharma et al. in the CASE dataset paper and shows a subject using the JERI annotation device while watching an emotion stimulus [1, 96].

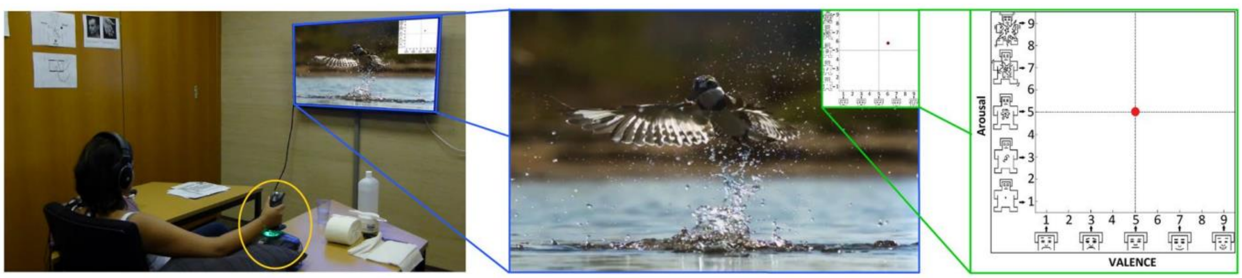


Fig. 14. Image from CASE Dataset Paper showing JERI Device for Continuous Arousal and Valence Annotation. Figure from [1] by Karam Sharma is licensed under CC BY 4.0.

The CASE dataset consists of physiological data from 30 subjects as they watched multiple videos designed to elicit certain emotions. The physiological data recorded include electroencephalogram (ECG) in milliVolt, blood volume pulse (BVP) in percentage, galvanic skin response (GSR) in microSiemens, skin temperature in degree Celsius, electromyography (EMG) for the zygomaticus major, corrugator supercilii, and trapezius muscles in mV, and respiration rate. The emotion eliciting videos were meant to elicit amused, bored, neutral, relaxed, and scared discrete emotions. Blank blue screens were shown before and after each viewing session to gather data for a baseline of the subject.

Due to the unique nature of this dataset, the arousal, valence, and discrete emotion labels are continuous and real-time. This allows the possibility of creating regression models which accurately predict arousal and valence using features generated from the physiological data as the independent variables. These regression models enable the creation of arousal and valence predictions from physiological data. The predicted arousal and valence can then be used along with the original physiological features to train classifiers on the discrete emotion labels of the CASE dataset. This method is enabled by the continuous and coupled nature of the arousal and valence labels along with the physiological data and discrete emotion labels in the CASE dataset.

3.2 Labels Preprocessing

The labels of the CASE dataset consist of the individual video labels consisting of an integer value for each video and the continuous arousal and valence data gathered using the JERI joystick from the subjects while they watched the videos. The impact of these preprocessing steps on the overall accuracy of the classification is shown in the Results section.

3.2.1 Video Label Relabeling and Resampling

Each video shown to the subjects in the CASE dataset was meant to elicit a certain emotion within the subjects. The elicited emotions were amused, bored, neutral, relaxed, and scared. The integer video labels were converted to emotion labels by grouping data from each video for a particular emotion together into the same label. Table 7 below shows the conversion from video labels to emotion labels.

TABLE 7. CONVERSION FROM VIDEO TO EMOTION LABELS FROM CASE DATASET

Video Label	Emotion Label	Emotion
10	0	Neutral
1	1	Amused
2	1	Amused
3	2	Bored
4	2	Bored
5	3	Relaxed
6	3	Relaxed
7	4	Scared
8	4	Scared

3.2.2 Arousal and Valence Resampling

The arousal and valence data were sampled at 20 Hz and the physiological data was sampled at 1000 Hz in the CASE dataset. Because of this, the arousal and valence were resampled to 1000 Hz using a

Fourier-domain moving window. Fig. 15 and Fig. 16 below show the resampled arousal and valence values compared to the original values.

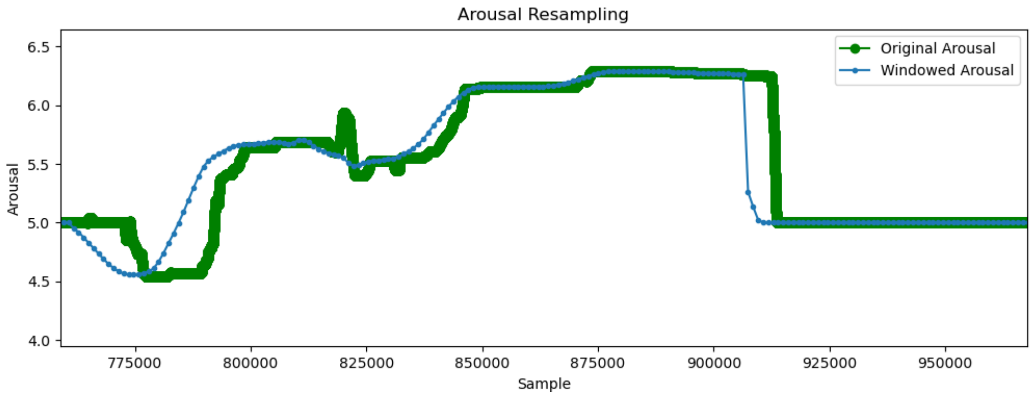


Fig. 15. Zoomed in View of Resampled versus Original Arousal Labels.

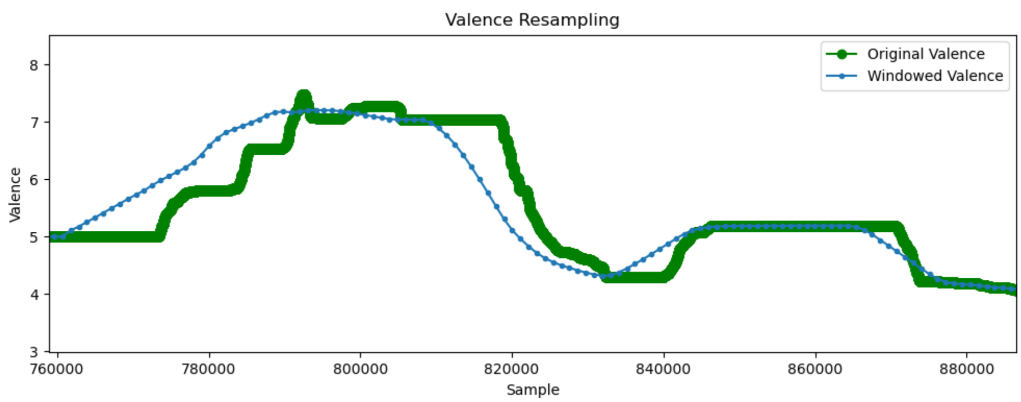


Fig. 16. Zoomed in View of Resampled versus Original Valence Labels.

3.2.3 Arousal and Valence Reordering

The original order of the arousal and valence values are arbitrary based on the order of the emotion-eliciting videos shown to the CASE subjects during the experiments. The arousal labels were thus reordered by emotion class so that the data is in lowest to highest arousal in the order: *bored, relaxed, neutral, amused, scared*. The valence labels were also reordered so that the labels are from negative to positive valence: *scared, bored, neutral, relaxed, amused*. The reordering allows the regression model to fit the arousal and valence more easily since they are ordered from smallest to largest values respectively. The physiological features were also reordered to match the arousal ordering and valence ordering respectively so that they could be used as independent input variables to the regression arousal and valence regression models. After the regression is fit and the predicted arousal and valence values are created, these values are then ordered back into the original order of the datapoints in order to concatenate them to the original physiological feature set for input into the classification models. Fig. 17 and Fig. 18 show the reordering of the preprocessed arousal and valence values annotated by the CASE subjects using the JERI device in the CASE dataset.

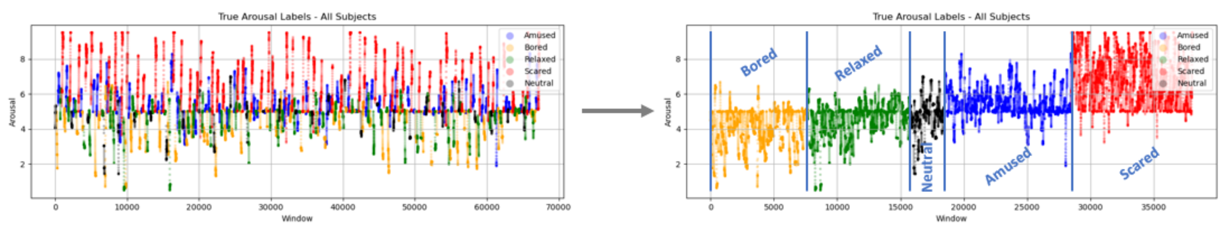


Fig. 17. Reordering of Arousal Labels from Lowest to Highest Arousal.

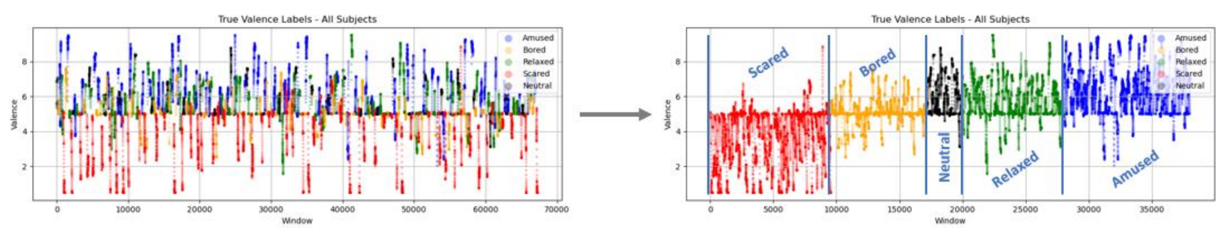


Fig. 18. Reordering of Valence Labels from Negative to Positive Valence.

3.2.4 Arousal and Valence Gaussian Distribution

A Gaussian distribution with the same mean and a fourth of the standard deviation of the arousal and valence labels was created to replace the original arousal and valence labels, respectively. This was done to artificially increase the separation between discrete emotion classes within the arousal and valence labels since the original arousal and valence were self-annotated by the CASE subjects. By human nature, the self-annotation introduces noise into the arousal and valence ground truth values, so this method of converting the original annotated values into Gaussian distributions slightly increases the separation between classes as shown in Fig. 19 below. The results of this preprocessing method are shown in Fig. 62 in the Results section. Fig. 19 shows the progression of the preprocessing of the self-annotated arousal and valence values in the CASE dataset into the datapoints used for regression model training.

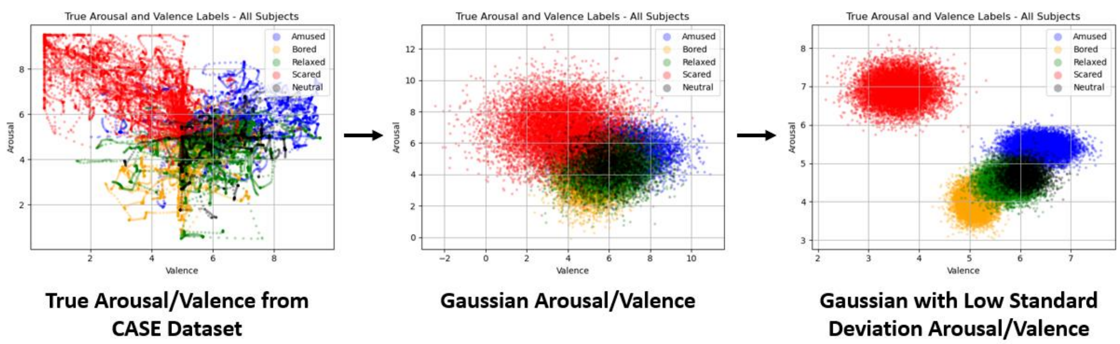


Fig. 19. True Arousal and Valence Preprocessing Steps before Inputting into Regression Models.

3.2.5 Arousal and Valence Scaling

Finally, the arousal and valence values were centered around zero and scaled from negative one to positive one to match the scaling of the physiological features. Fig. 20 below shows the scaling of the arousal and valence.

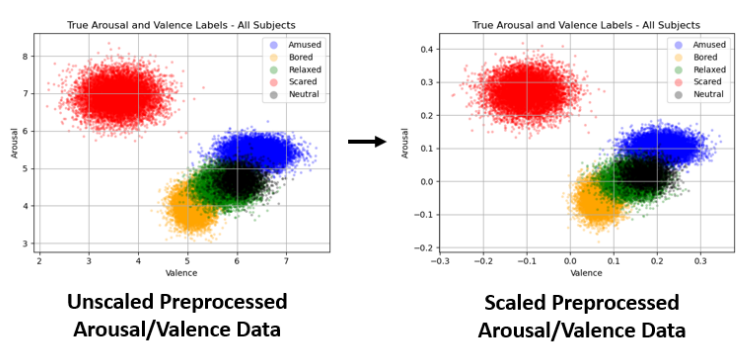


Fig. 20. Arousal and Valence Feature Rescaling.

3.3 Physiological Data Preprocessing

The physiological data from the CASE dataset used in this work include BVP, GSR, and skin temperature. These signals were each filtered to remove high and low-frequency noise for feature extraction in the next step. A 3rd-Order Butterworth zero-phase shift forward-backward filter was used to prevent a phase shift in the resulting filtered signal. The BVP signal was bandpass filtered between 0.25-3 Hz, and the GSR and skin temperature signals were lowpass filtered at 1.5 Hz. Fig. 21, Fig. 22, and Fig. 23 below compare the unfiltered and filtered physiological data using the filtering methods described.

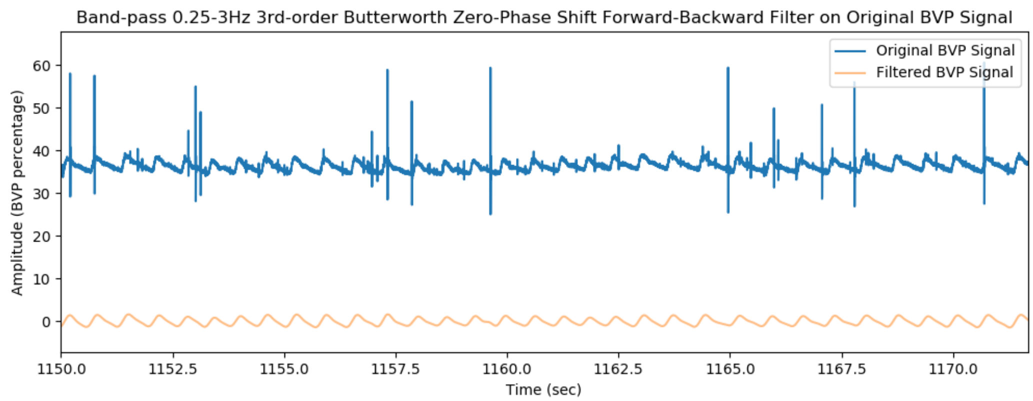


Fig. 21. BVP Signal Preprocessing.

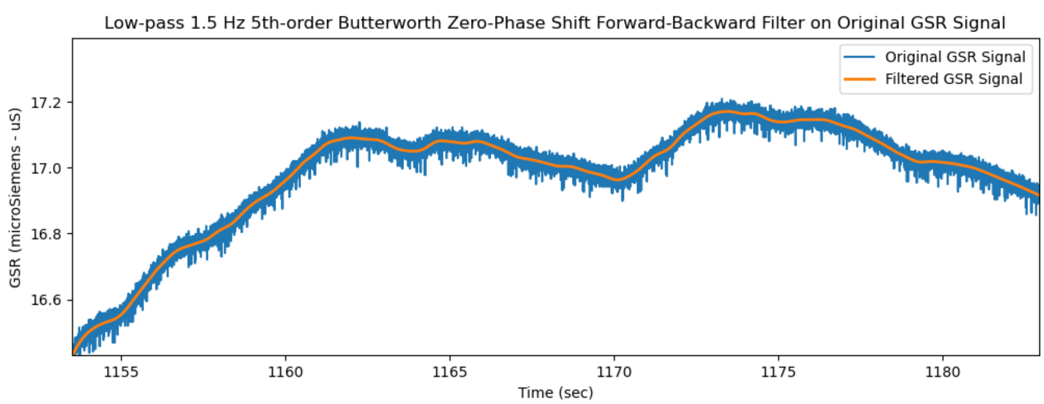


Fig. 22. GSR Signal Preprocessing.

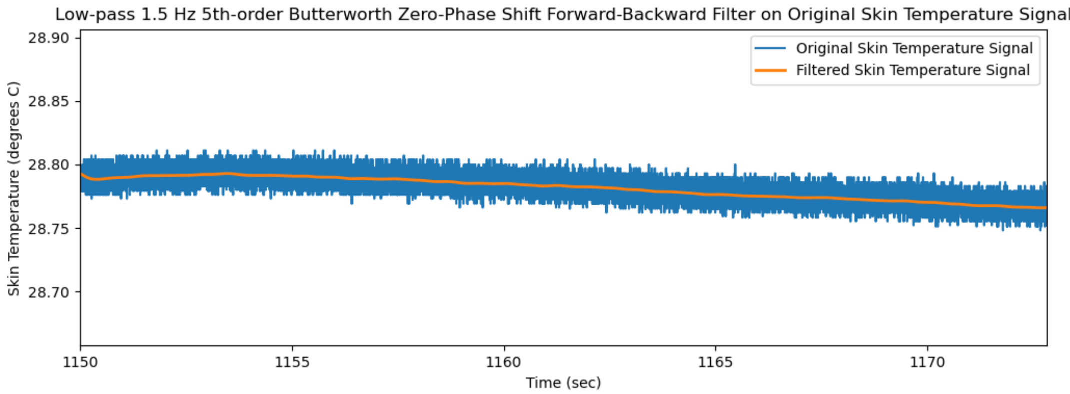


Fig. 23. Skin Temperature Signal Preprocessing.

3.4 Feature Extraction

Seventeen features were extracted from the physiological data of the CASE dataset to use as input data for the regression and classification models. These features represent physical responses to the emotional stimuli in the videos shown to the CASE subjects. These features were extracted from three physiological signals: thirteen from BVP, two from GSR, and two from skin temperature windowed in ten-second moving windows with one-second stride.

3.4.1 Windowing

Features were extracted from the filtered BVP, GSR, and skin temperature signals using a ten-second moving window with a window stride or overlap of one second. Fig. 24 below shows the moving windows of data that are taken from the original dataset and used to create time-correlated feature vectors.

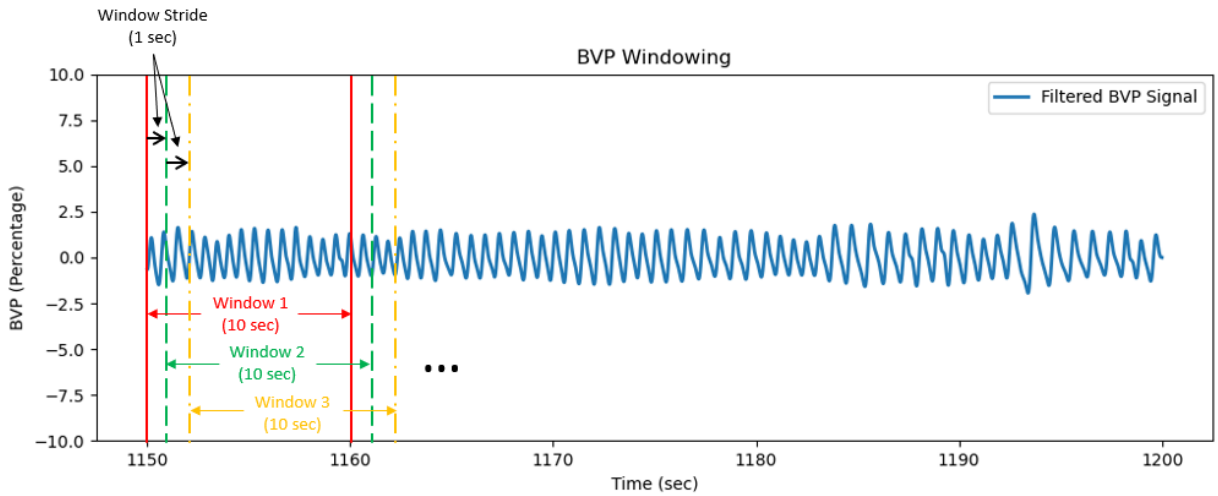


Fig. 24. Feature Extraction Windowing.

3.4.2 BVP Features

Blood Volume Pulse (BVP) is a measurement technique of heart function by using a photo-detector pair placed on the skin to detect the change in amplitude of infrared light reflections due to changes in the volume of blood as it circulates through blood vessels. It is a periodic signal representing the beating of a heart and is extensively used in medicine and recreation for gathering information about a heart's health and function. To extract features from the BVP signal, an open-source python library was used: Heartpy [97]. This library extracts the following features from BVP by using peak detection to determine when the heart is beating from each signal as shown in Table 8 below. Fig. 25 shows the peak detection of the heartpy Python library on the BVP physiological data [97].

TABLE 8. ECG AND BVP FEATURES EXTRACTED USING HEARTPY LIBRARY

Abbreviation	Name	Description
BPM	Beats Per Minute	Heart rate
IBI	Interbeat Interval	Variability of heart rate
SDNN	Standard Deviation of RR Intervals	Variability of R-R Intervals
SDSD	Standard Deviation of Successive Differences	Variability of differences between adjacent N-N distances in a time series.
RMSSD	Root Mean Square of Successive Differences of Intervals	Normalized differences between adjacent N-N distances in a time series.
pNN20	Proportion of Successive Differences above 20ms	Ratio of differences between adjacent N-N distances in a time series above 20ms
pNN50	Proportion of Successive Differences above 50ms	Ratio of differences between adjacent N-N distances in a time series above 50ms

TABLE 8. CONTINUED

Abbreviation	Name	Description
MAD	Median Absolute Deviation of RR Intervals	Variability of R-R Intervals
SD1	Standard Deviation Perpendicular to the Line of Identity in Poincaré Ellipse [98]	Related to fast changes of heartbeats in data (high-frequency spectrum)
SD2	Standard Deviation Parallel to the Line of Identity in Poincaré Ellipse [98]	Long-term variations of R-R interval (low-frequency spectrum)
S	Area of Poincaré Ellipse [98]	Aggregate measure of low and high-frequency heart rate information
SD1/SD2	Ratio of Poincaré Standard Deviations [98]	Ratio of short and long variations in R-R interval
Inferred Breathing Rate	Estimated Breathing Rate based on Heart Rate	Estimation of respiration rate from heart rate as defined in [99]

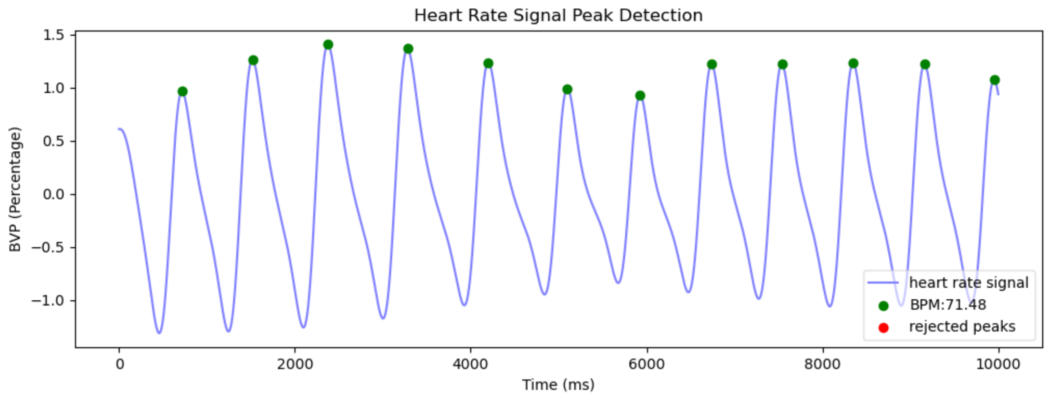


Fig. 25. BVP Heart Rate Peak Detection for BVP Feature Extraction.

3.4.3 GSR Features

Two features were extracted from GSR: the average value within the window, and the slope of the signal as seen in Fig. 26. This represents the magnitude of the GSR signal and how much the signal is increasing or decreasing within the window.

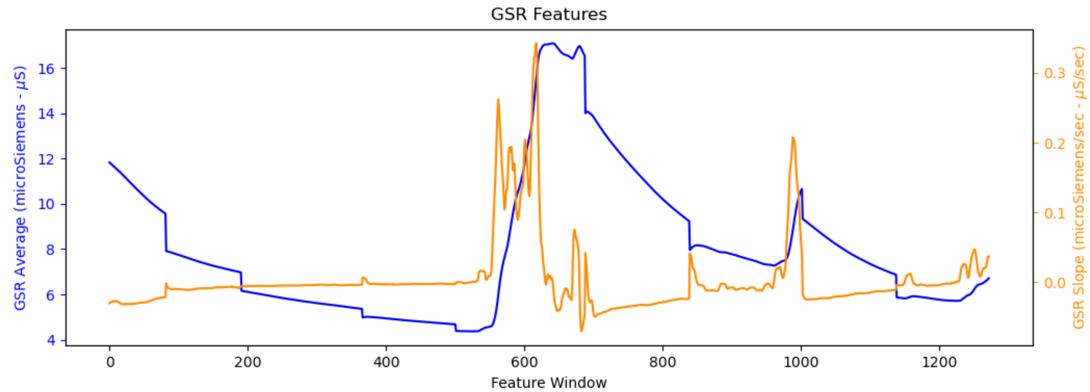


Fig. 26. Features Extracted from GSR.

3.4.4 Skin Temperature Features

The skin temperature features represent the magnitude and rate of change with the average and slope of the signal within the 10-second window as seen in Fig. 27.

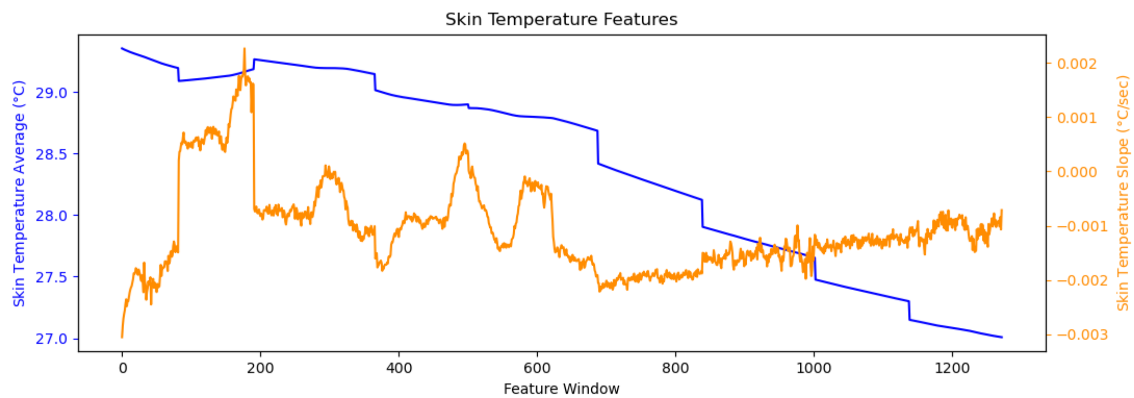


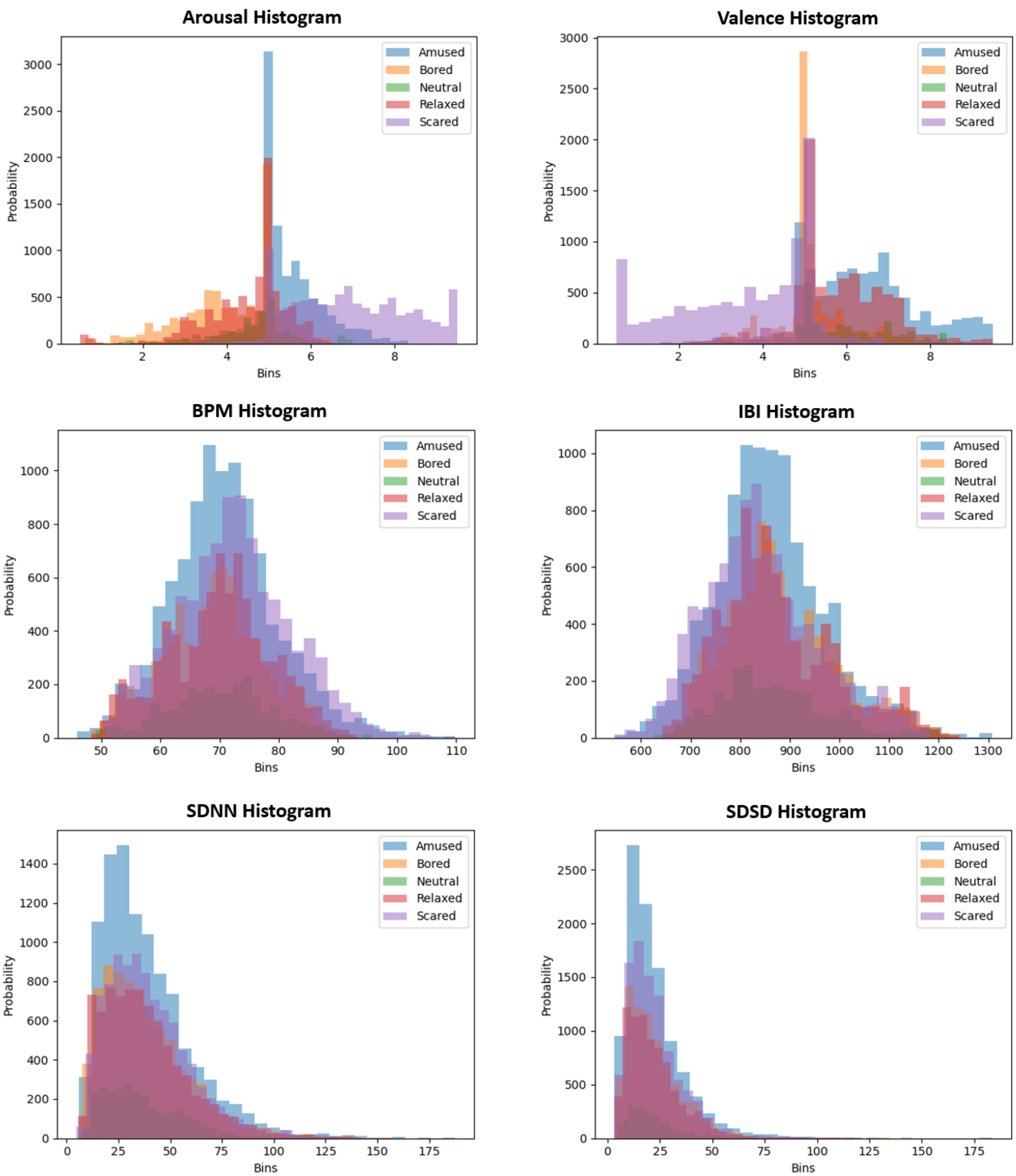
Fig. 27. Features Extracted from Skin Temperature.

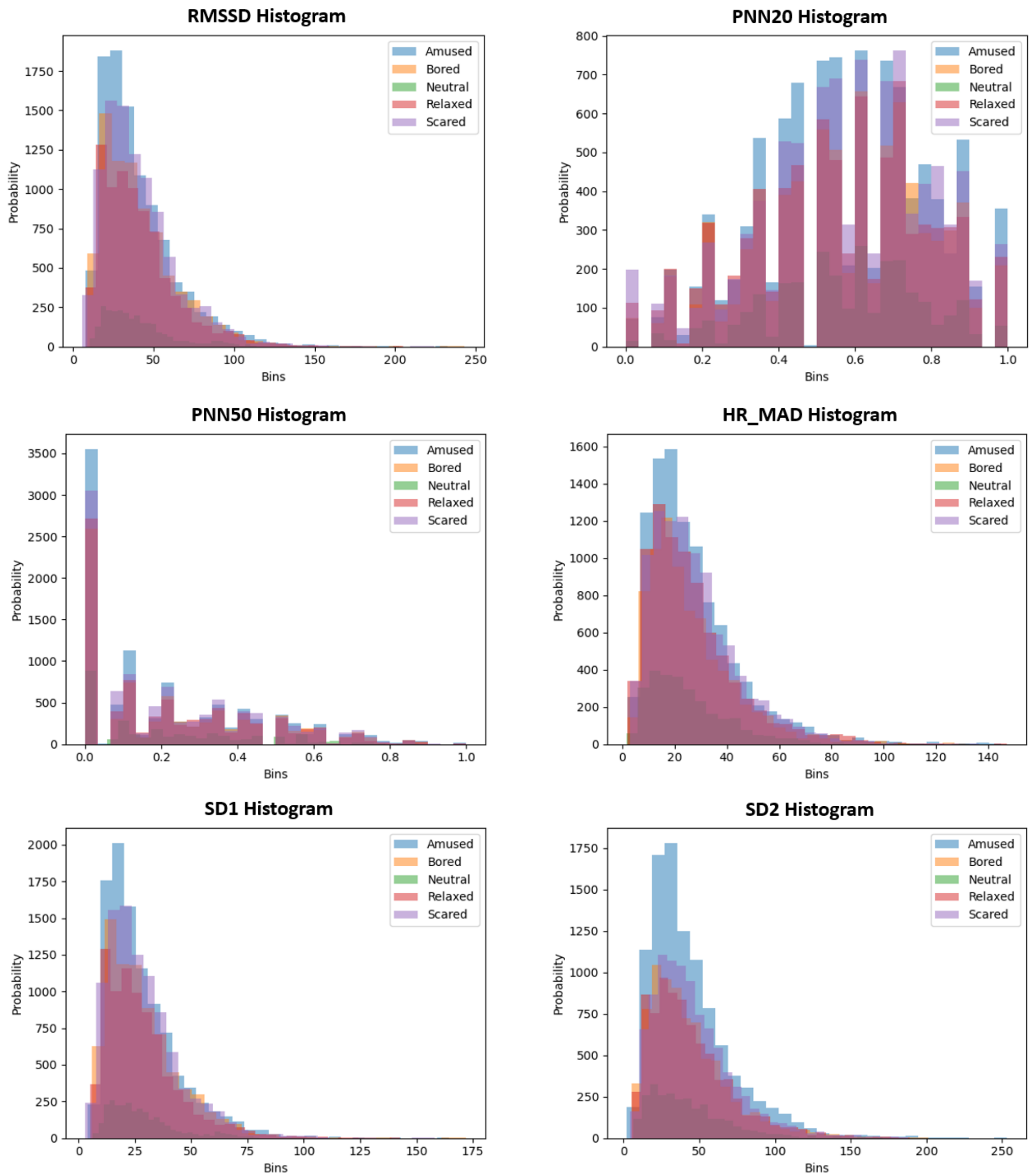
3.5 Feature Set Statistical Analysis

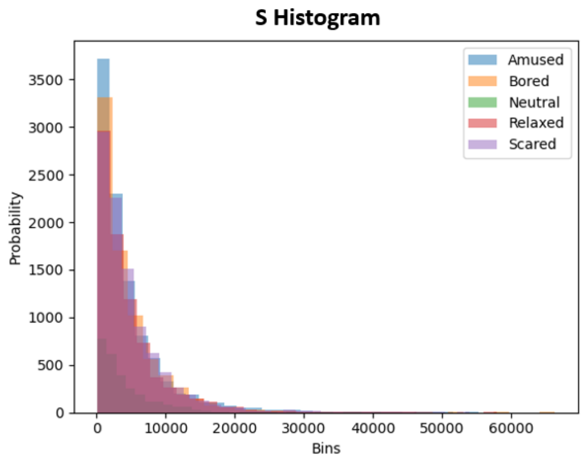
3.5.1 Feature Distribution

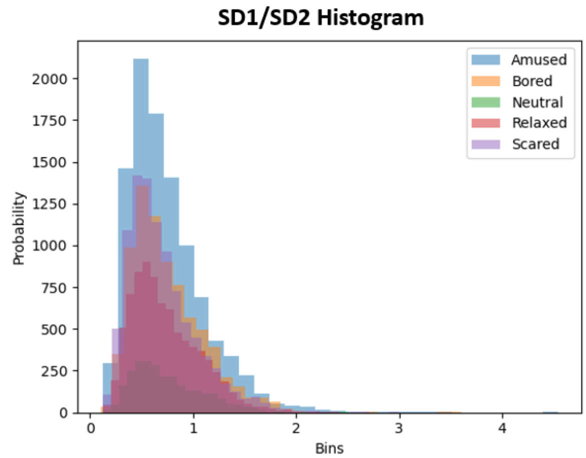
To gain greater insight into the behavior of the features generated, the distribution of each feature was plotted for all 30 subjects. This was done to determine what kind of distribution each feature had such as

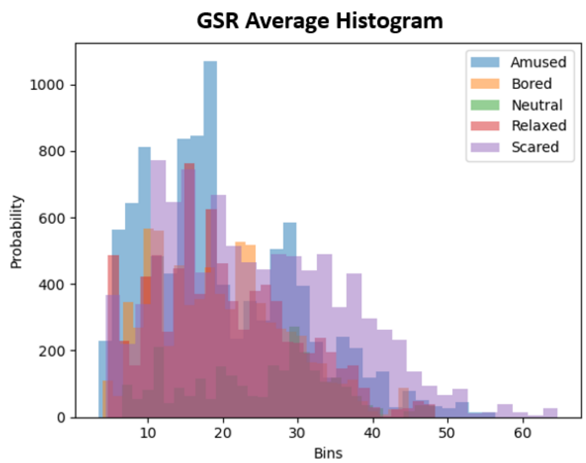
Gaussian, Poisson, single-modal, bimodal, multi-modal, or another type. Fig. 28 shows that most features follow a Gaussian and Poisson distribution while breathing rate follows a multi-modal distribution.

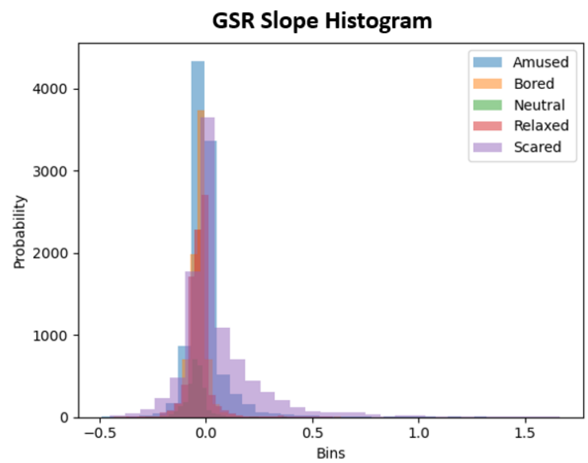


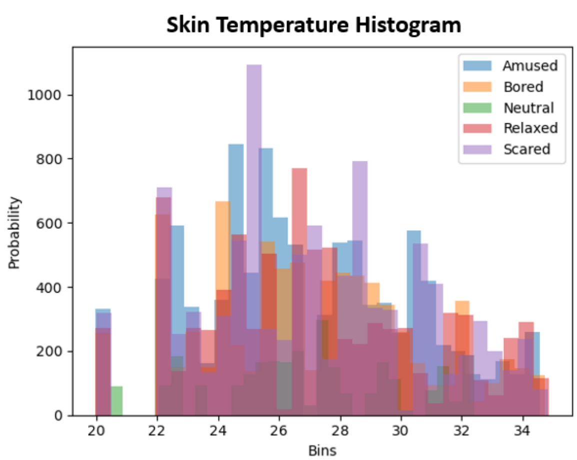

RMSSD Histogram
PNN20 Histogram
PNN50 Histogram
HR_MAD Histogram
SD1 Histogram
SD2 Histogram
Amused
Bored
Neutral
Relaxed
Scared
Probability
Bins

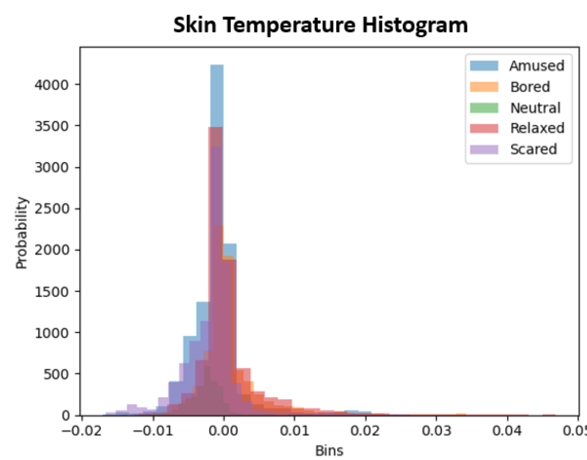

S Histogram
Amused
Bored
Neutral
Relaxed
Scared
Probability
Bins


SD1/SD2 Histogram
Amused
Bored
Neutral
Relaxed
Scared
Probability
Bins


GSR Average Histogram
Amused
Bored
Neutral
Relaxed
Scared
Probability
Bins


GSR Slope Histogram
Amused
Bored
Neutral
Relaxed
Scared
Probability
Bins


Skin Temperature Histogram
Amused
Bored
Neutral
Relaxed
Scared
Probability
Bins


Skin Temperature Histogram
Amused
Bored
Neutral
Relaxed
Scared
Probability
Bins

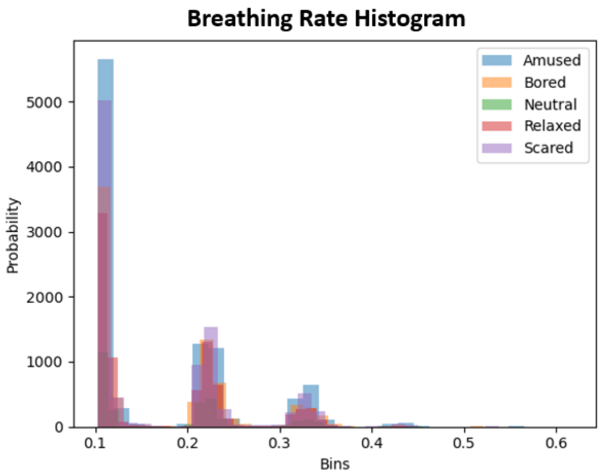


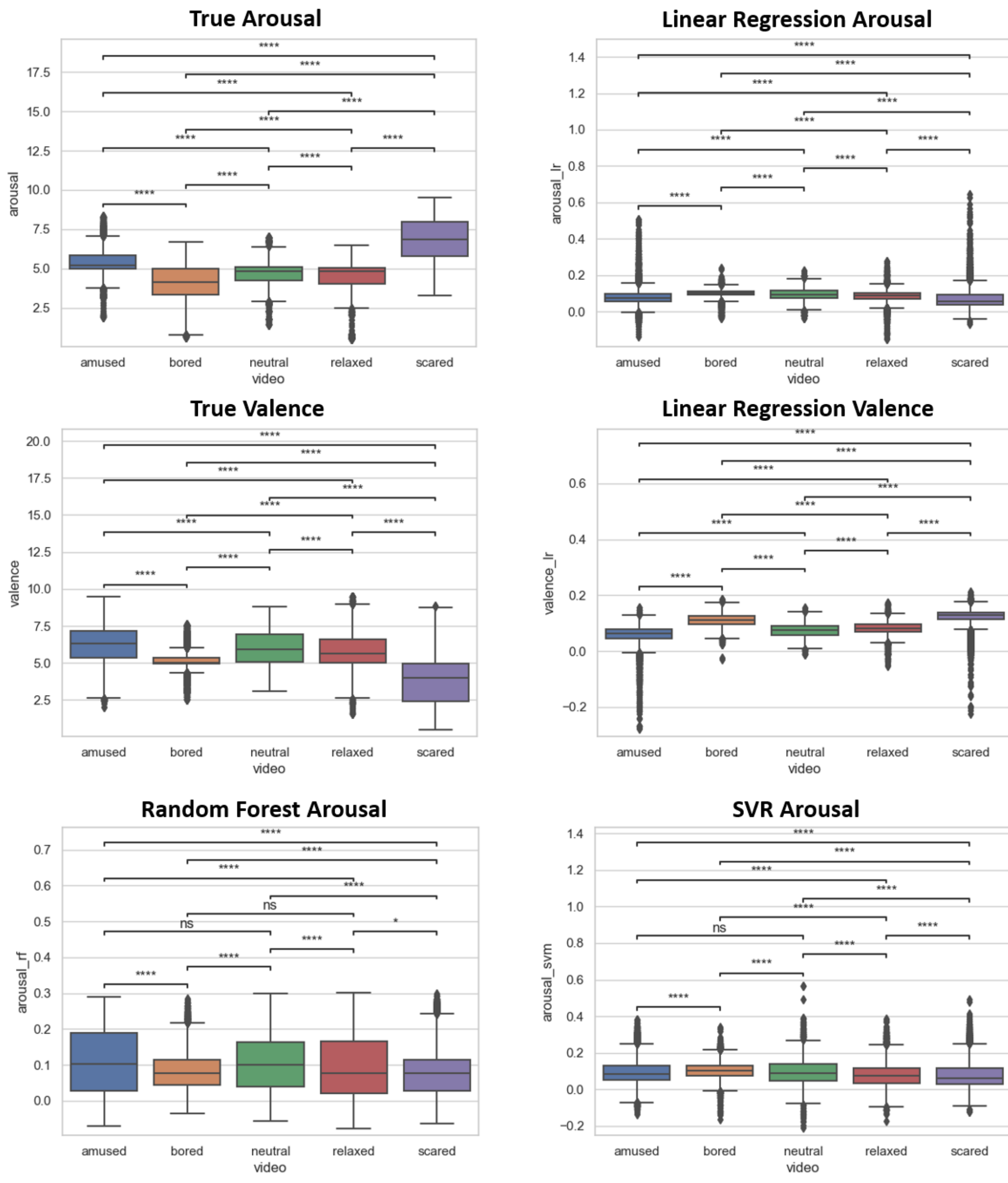
Fig. 28. Feature Distributions of ECG, BVP, GSR, and Skin Temperature Features before Oversampling.

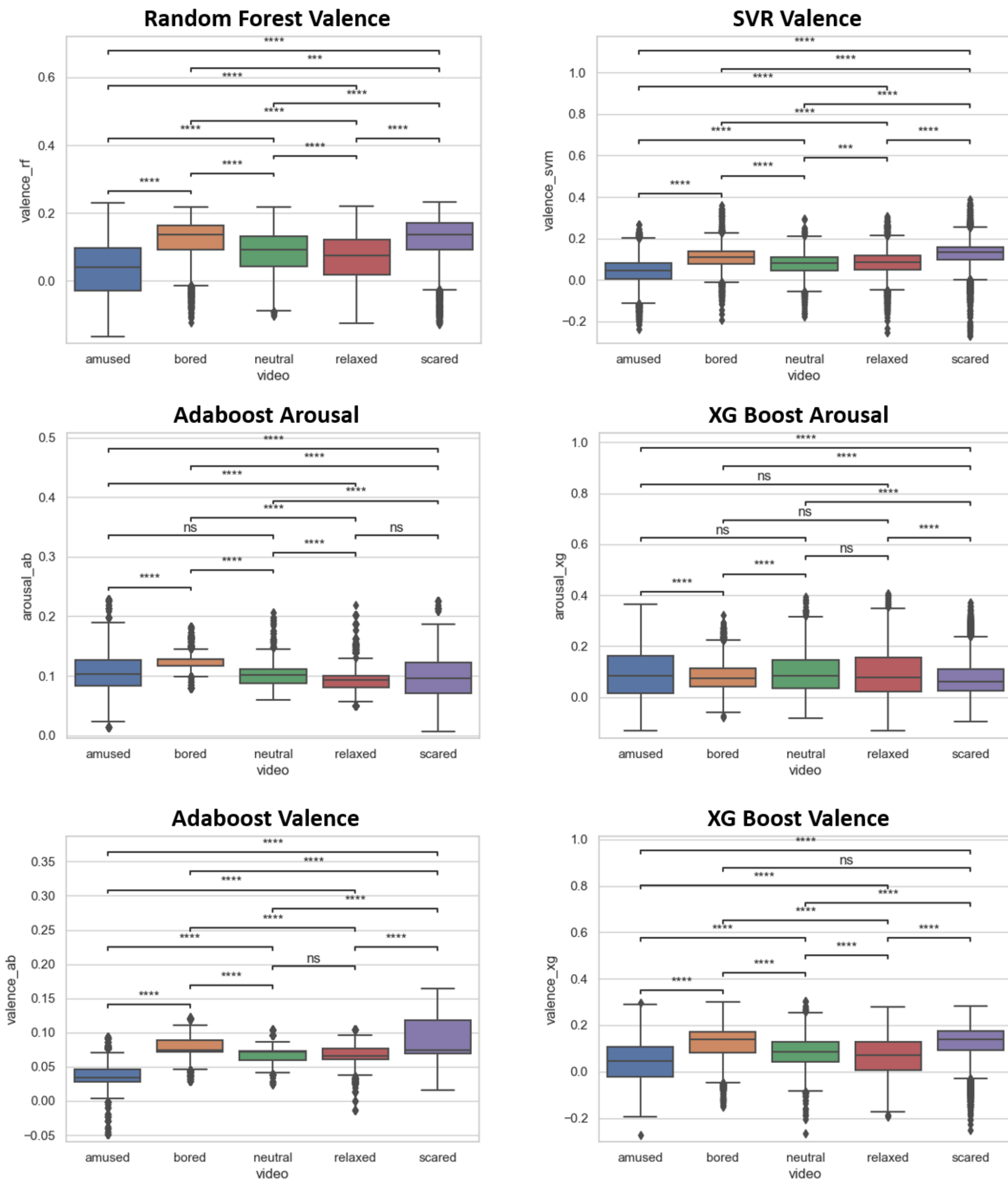
3.5.2 Statistical Significance

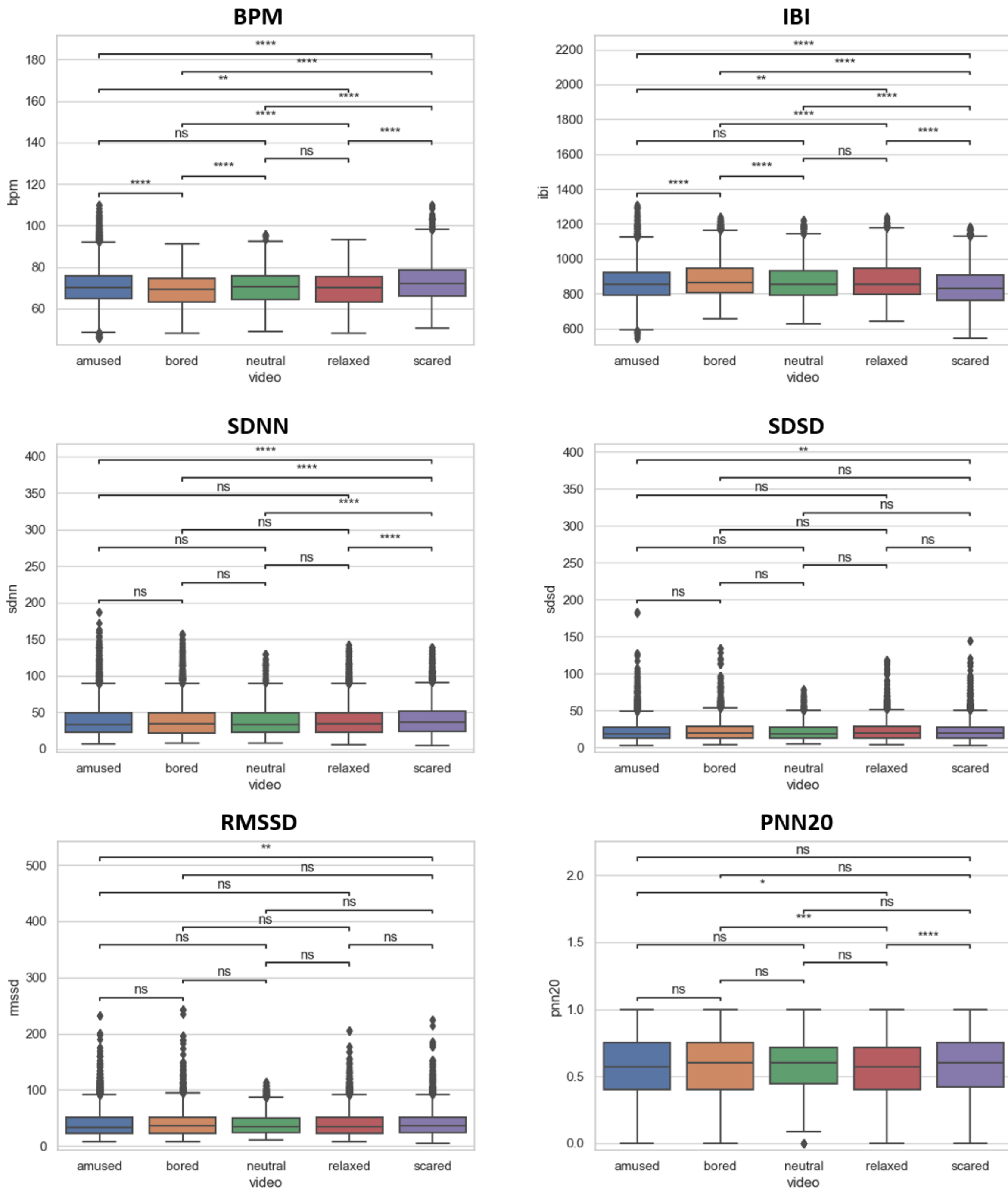
A Mann-Whitney U test was used to determine the statistical significance of every emotion class compared to every other emotion class within that specific feature. In order to calculate the p-value and generate graphs visualizing the statistical significance between classes, the "seaborn" and "statannot" Python libraries were used [100, 101]. The p-value between each class is displayed on the graph through stars where more stars symbolize a lower p-value and thus a greater statistical significance according to Table 9 below. Fig. 29 graphically shows the statistical significance of the difference in each emotion class using each feature. Some features have statistically significant differences when comparing every class with one another meaning they are very information-rich, while other features have multiple classes which show no statistically significant difference meaning they are relatively information-poor. The symbols noted in Table 9 are shown in Fig. 29 indicating each classes' p-value from one another using information from each feature.

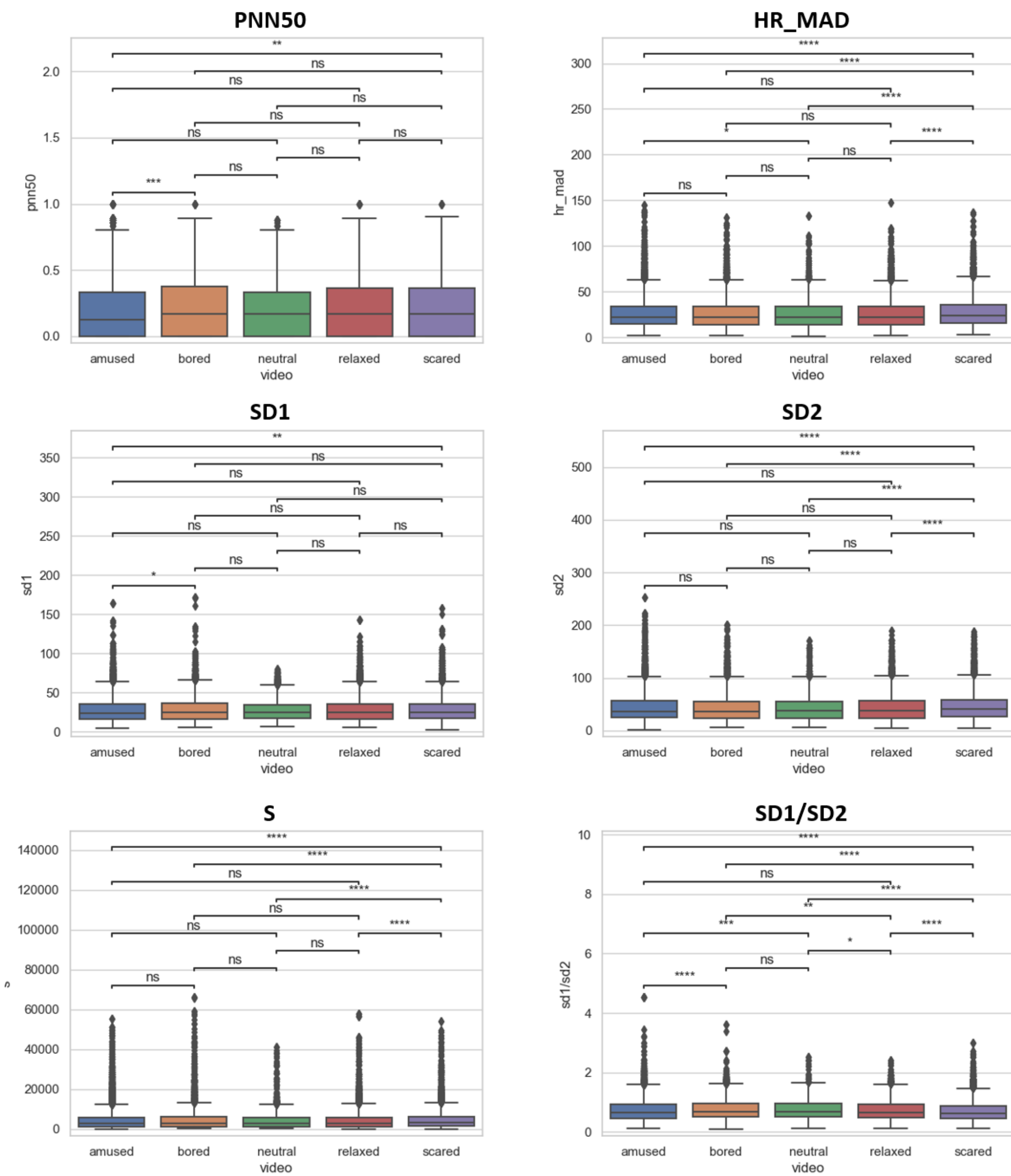
TABLE 9. P-VALUE ANNOTATION LEGEND

Symbol	P-Value Range
ns (no significance)	5.00e-02 < p <= 1.00e+00
*	1.00e-02 < p <= 5.00e-02
**	1.00e-03 < p <= 1.00e-02
***	1.00e-04 < p <= 1.00e-03
****	p <= 1.00e-04


True Arousal
arousal
amused bored neutral relaxed scared
video
Linear Regression Arousal
arousal_lr
amused bored neutral relaxed scared
video
True Valence
valence
amused bored neutral relaxed scared
video
Linear Regression Valence
valence_lr
amused bored neutral relaxed scared
video
Random Forest Arousal
arousal_rf
amused bored neutral relaxed scared
video
SVR Arousal
arousal_svm
amused bored neutral relaxed scared
video


Random Forest Valence
valence_rf
SVR Valence
valence_svm
Adaboost Arousal
arousal_ab
XG Boost Arousal
arousal_xg
Adaboost Valence
valence_ab
XG Boost Valence
valence_xg
amused
bored
neutral
relaxed
scared
video
ns
****
***


BPM
bpm
IBI
ibi
SDNN
sdnn
SDSD
sdsd
RMSSD
rmssd
PNN20
pnn20
amused
bored
neutral
relaxed
scared
video


PNN50
pnn50
HR_MAD
hr_mad
SD1
sd1
SD2
sd2
S
s
SD1/SD2
sd1/sd2
amused
bored
neutral
relaxed
scared
video

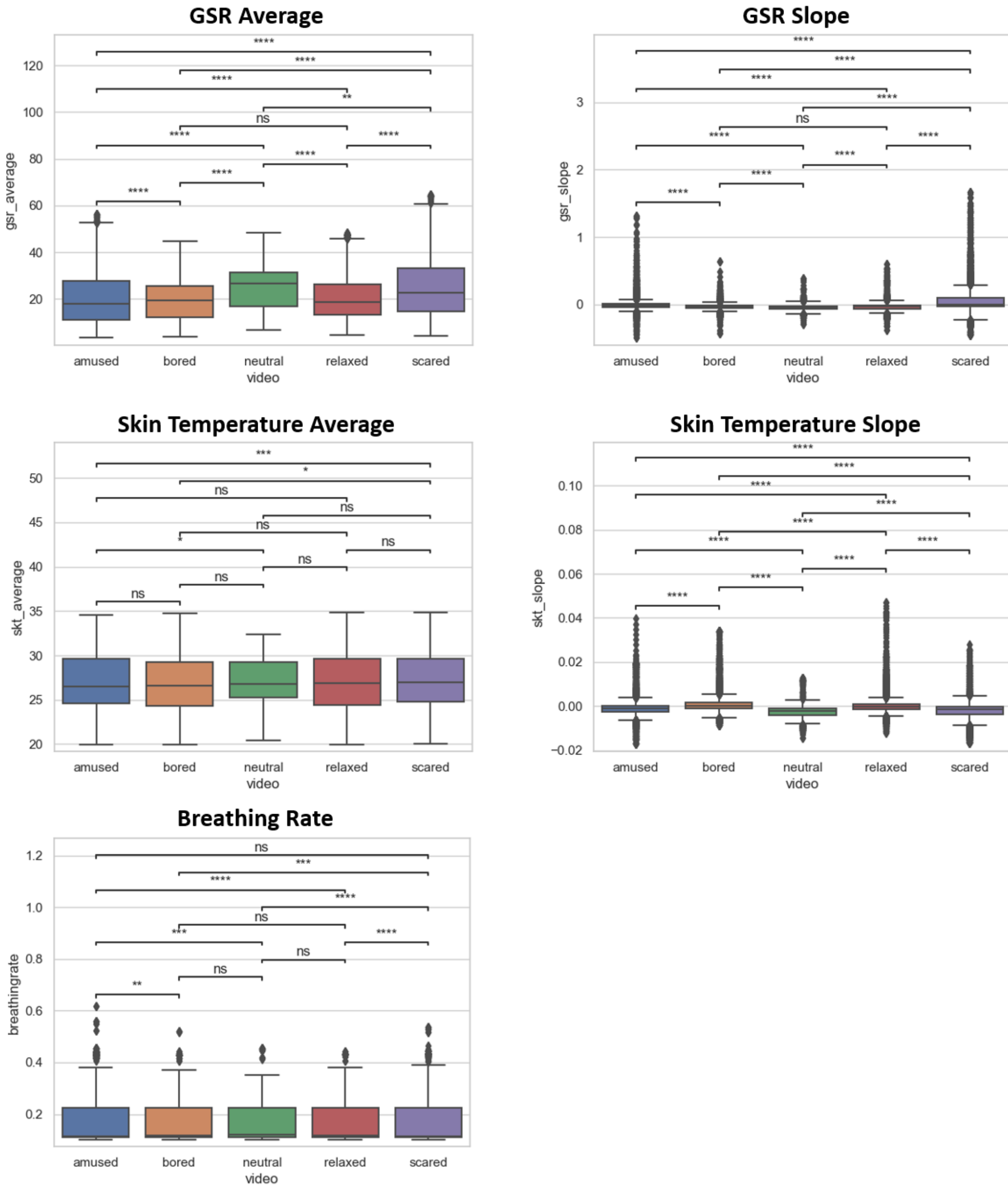


Fig. 29. Mann-Whitney U-Test Statistical Significance Between Emotion Classes for Each Feature.

3.6 Feature Postprocessing

3.6.1 Imbalanced Class Oversampling

Since the datasets for each discrete emotion class were of different lengths in time, there were more feature vectors for some classes than for others. This was especially apparent in the "neutral" class which had only one video that the subjects watched instead of the two videos for the other emotion classes. Due to this, an oversampling method was implemented using the Borderline-SMOTE SVM method [102]. Using this method, feature vectors were generated so that each class contained the same number of feature vectors for training the classification model. This prevented any class imbalance issues commonly seen in classification models trained on highly imbalanced classes. Table 10 gives an idea of the class imbalance before and after Borderline-SMOTE SVM oversampling, and Fig. 30 graphically shows the results of oversampling on two example features: arousal and valence.

TABLE 10. BORDERLINE-SMOTE SVM OVERSAMPLING RESULTS

Emotion	Duration (sec) (per subject)	# of Feature Vectors (per subject)	# of Feature Vectors (30 subjects)	# of Feature Vectors Oversampled
Amusing	358.7	348	10017	10017
Boring	278.8	268	7673	10017
Relaxed	291.9	281	8073	10017
Scary	340.8	330	9560	10017
Neutral	101.5	91	2735	10017

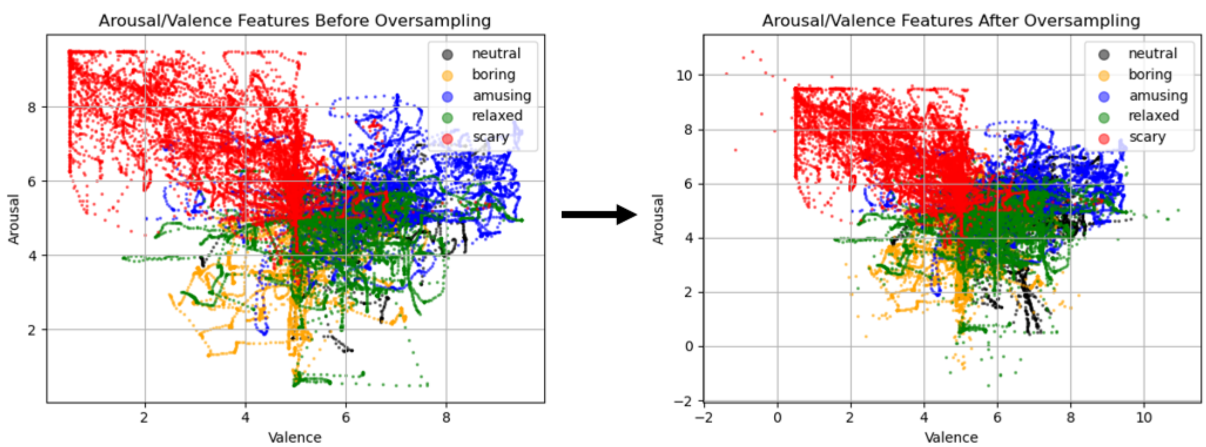


Fig. 30. Example Feature Oversampling of Arousal and Valence Features using the Borderline-SMOTE SVM method.

3.6.2 Standardization

Feature standardization serves multiple purposes. It scales the features all to the same range so that some models, such as the neural network, treat each feature equally in terms of affecting weights and errors during training. It also reduces the differences between subjects by normalizing the features to a set range.

Each feature is standardized across all subjects by removing the mean and scaling to unit variance using the equation:

$$z = \frac{x - \mu}{\sigma} \tag{8}$$

where z is the standardized feature value, x is the original feature value, μ is the mean of a feature, and σ is the standard deviation of the feature.

3.7 Arousal and Valence Regression Model Training

Multiple regression model methods were tested for creating arousal and valence prediction models. These regression models were used to create predicted arousal and valence values from the physiological features. One model was created for predicting arousal and another was created for predicting valence. The physiological features were used as independent variables and the arousal or valence labels from the CASE dataset were used as the dependent variable in training the models. The regression models tested were linear regression, random forest regressor, support vector regressor (SVR), AdaBoost, and XG Boost. These models were then used to predict arousal and valence from the physiological features. Doing this allowed the use of the physiological features to create two more features: predicted arousal and predicted valence. These new features can then be concatenated with the original physiological features to create a much more robust feature set for the classification models. Fig. 31 below shows the true arousal/valence values from the CASE dataset for each emotion class.

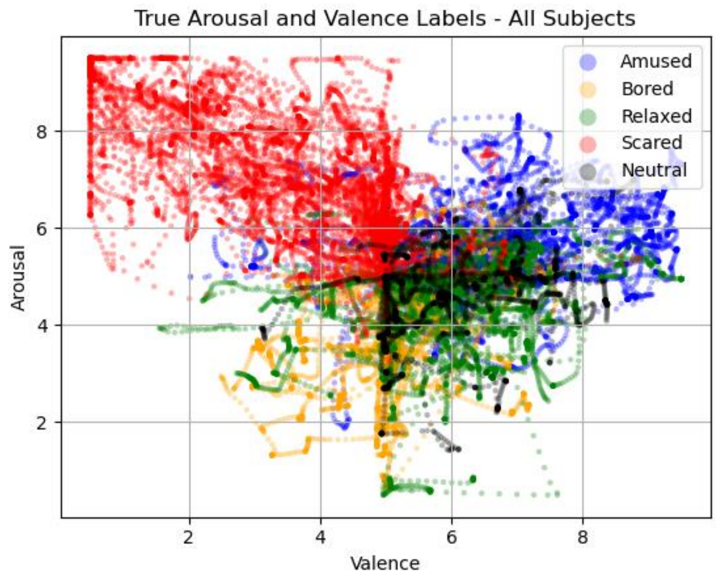


Fig. 31. True Arousal and Valence Features in 2D Circumplex Model Space.

Fig. 32 below shows the entire regression process using the self-annotated arousal and valence labels from the CASE subjects, the physiological features extracted using the methods above in the Feature Extraction section, and the regression models trained on the arousal and valence labels using the physiological features as the input. It shows the arousal and valence preprocessing steps of reordering and Gaussian conversion as described in the Labels Preprocessing section. There are two separate models: one for arousal and another for valence. Each model is trained using 5-fold cross-validation and the predicted arousal/valence values from each fold is saved as features for inputting into the final classification model explained in the next section Emotion Classification.

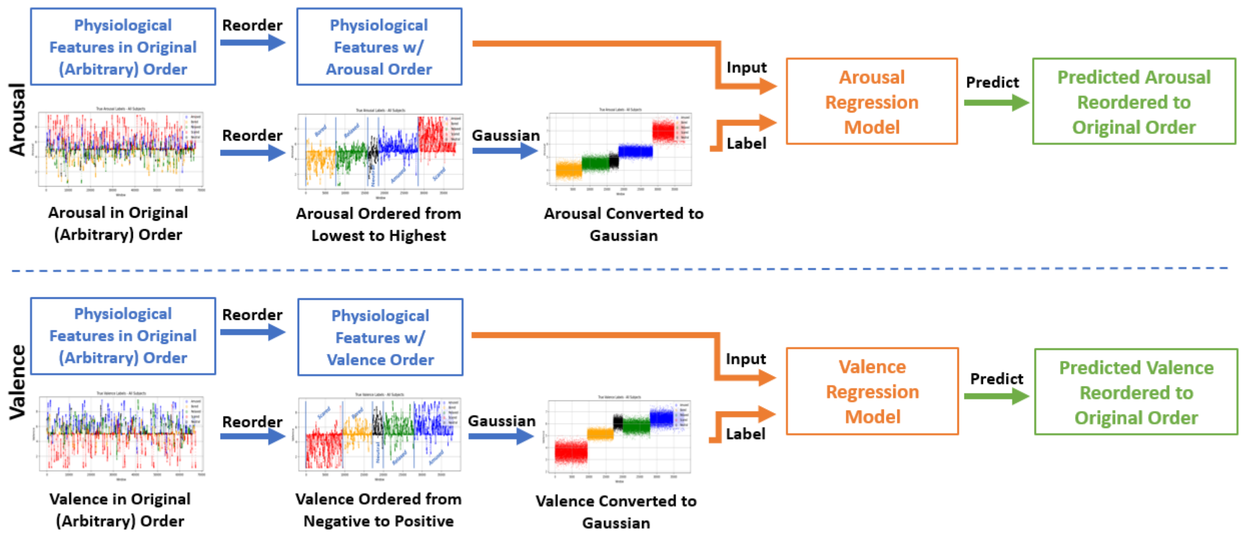


Fig. 32. Arousal and Valence Preprocessing and Regression Workflow

3.7.1 Linear Regression

A linear regression model was created using the Python sklearn library and mean absolute error, mean squared error, and root mean squared error metrics were used as evaluation metrics [103]. An ordinary least squares regressor is used which minimizes the residual sum of squares between the observed and predicted values [103]. As seen in the Feature Distribution section, the distribution of the input features closely represents a Gaussian distribution, and the features were also post-processed to remove noise and rescaled using standardization as described in the Feature Postprocessing section. These steps help improve the prediction reliability of linear regression models and their effectiveness at improving the accuracy of the ensemble model as a whole is shown in the Results section. Fig. 33 below graphs the predicted 2D arousal and valence space from the output of the linear regression model.

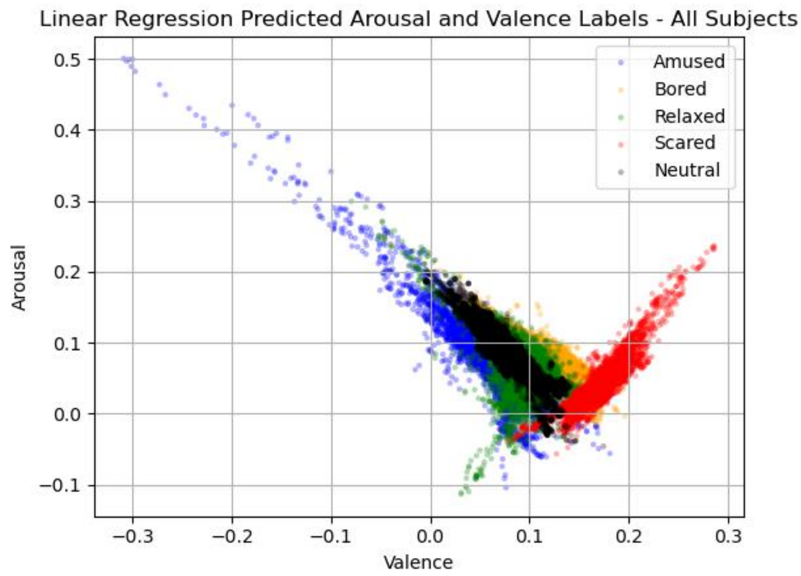


Fig. 33. Predicted Arousal and Valence Features from Linear Regression.

3.7.2 Random Forest Regressor

A random forest regression model was created using the Python sklearn library and mean absolute error, mean squared error, and root mean squared error metrics were used as evaluation metrics [103]. Table 11 below shows the parameters used in the model and Fig. 34 below shows the predicted 2D arousal and valence space.

TABLE 11. RANDOM FOREST REGRESSOR HYPERPARAMETER VALUES

Hyperparameter	Value
NUMBER OF ESTIMATORS (TREES)	100
CRITERION	'SQUARED ERROR'
MAX DEPTH	LESS THAN MIN SAMPLES SPLIT SAMPLES
MIN SAMPLES SPLIT	2

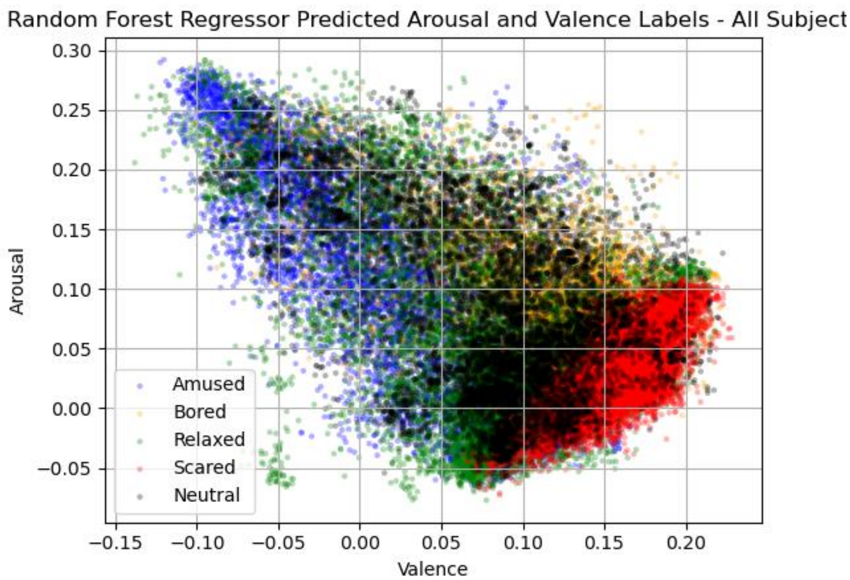


Fig. 34. Predicted Arousal and Valence Features from Random Forest Regressor.

3.7.3 Support Vector Regressor

A support vector regressor model was created using the Python sklearn library and mean absolute error, mean squared error, and root mean squared error metrics were used as evaluation metrics [103]. Table 12 below shows the parameters used in the model and Fig. 35 below shows the predicted 2D arousal and valence space.

TABLE 12. SVR HYPERPARAMETER VALUES

Hyperparameter	Value
KERNEL	RBF
DEGREE	3
GAMMA	SCALE
TOLERANCE	1E-3
C	1.0
EPSILON	0.1

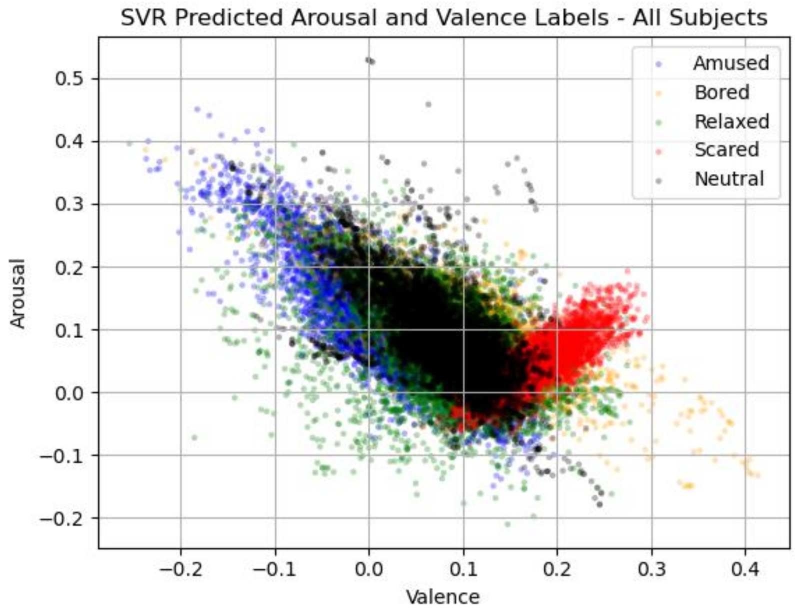


Fig. 35. Predicted Arousal and Valence Features from Support Vector Regressor.

3.7.4 Adaboost Regressor

An Adaboost model was created using the Python sklearn library and mean absolute error, mean squared error, and root mean squared error metrics were used as evaluation metrics [103]. Table 13 below shows the parameters used in the model and Fig. 36 below shows the predicted 2D arousal and valence space.

TABLE 13. ADABOOST HYPERPARAMETER VALUES

Hyperparameter	Value
BASE ESTIMATOR	DECISION TREE REGRESSOR
NUMBER OF ESTIMATORS (TREES)	50
LEARNING RATE	1.0
LOSS FUNCTION	LINEAR

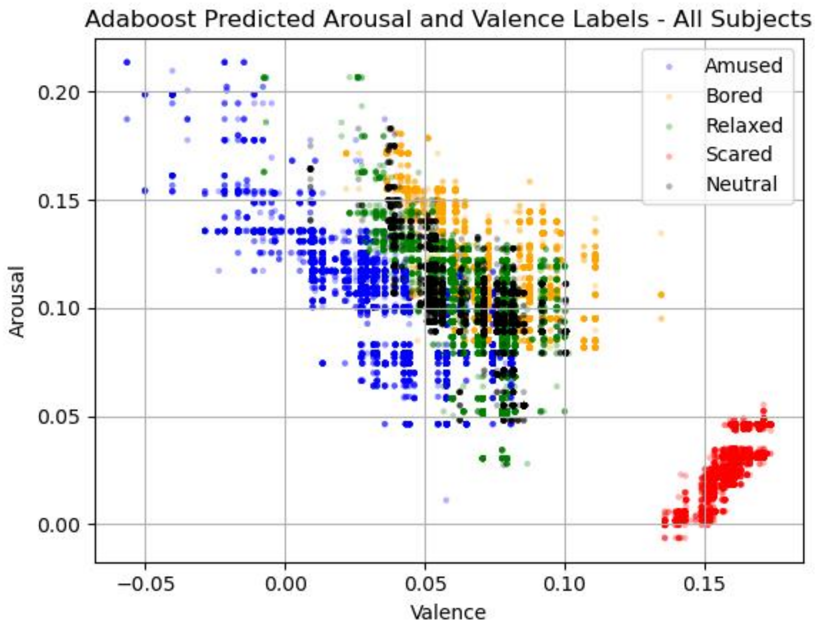


Fig. 36. Predicted Arousal and Valence Features from Adaboost Regressor.

3.7.5 XG Boost Regressor

An XG Boost model was created using the Python xgboost library and mean absolute error, mean squared error, and root mean squared error metrics were used as evaluation metrics [104]. Table 14 below shows the parameters used in the model and Fig. 37 below shows the predicted 2D arousal and valence space.

TABLE 14. XG BOOST HYPERPARAMETER VALUES

Hyperparameter	Value
BASE ESTIMATOR	GRADIENT BOOST TREE REGRESSOR
LEARNING RATE	0.3
MIN SPLIT LOSS	0
MAX DEPTH	6
MIN CHILD WEIGHT	1
MAX DELTA STEP	0
SUBSAMPLE	1
SAMPLING METHOD	"UNIFORM"

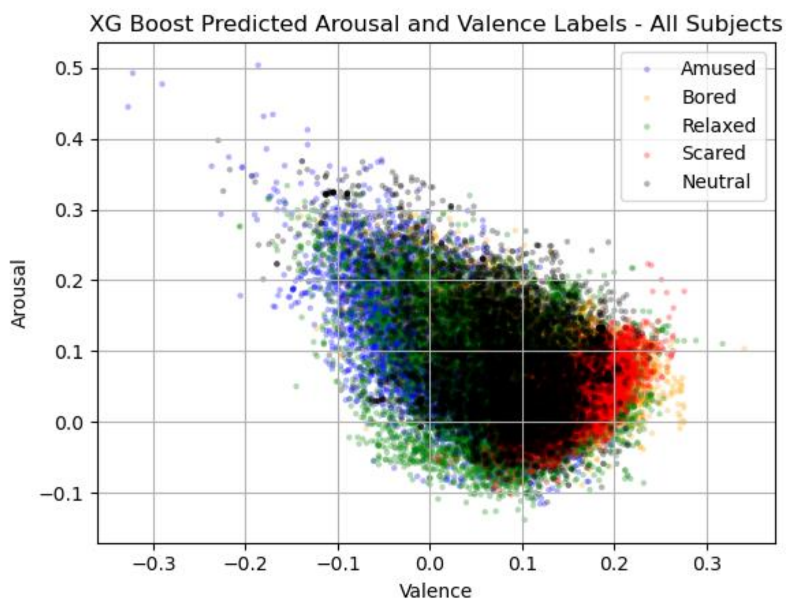


Fig. 37. Predicted Arousal and Valence Features from XG Boost Regressor.

3.8 Emotion Classification

The predicted arousal and valence labels from the regression models were then concatenated with the original physiological features to create the full feature set used as the input to train the classification models. Classification models were also trained with the “true” arousal and valence labels from the CASE dataset concatenated with physiological features as input features as well as trained on just the physiological features as a control comparison with this new method. The classification models tested were shallow neural network, 1D convolutional neural network (CNN) based on the “Alexnet” 2D CNN, LSTM, hyperparameter-tuned random forest, and hyperparameter-tuned support vector machine. Each classification model was tested with predicted arousal and valence features from each regression model described in the previous section.

Five-fold cross-validation was performed for each model, and the average and standard deviation of accuracy, AUC, F1 score, recall, and precision across the five folds are reported in the results section. The input features were not shuffled before the training and testing splits were created so that the data in the testing set contained data from subjects that the model had never seen before in training. This gives a better indication of the model’s performance and generalizability with data from subjects that it had never seen before. Fig. 38 below shows the training and testing splits used in the five-fold cross-validation model training.

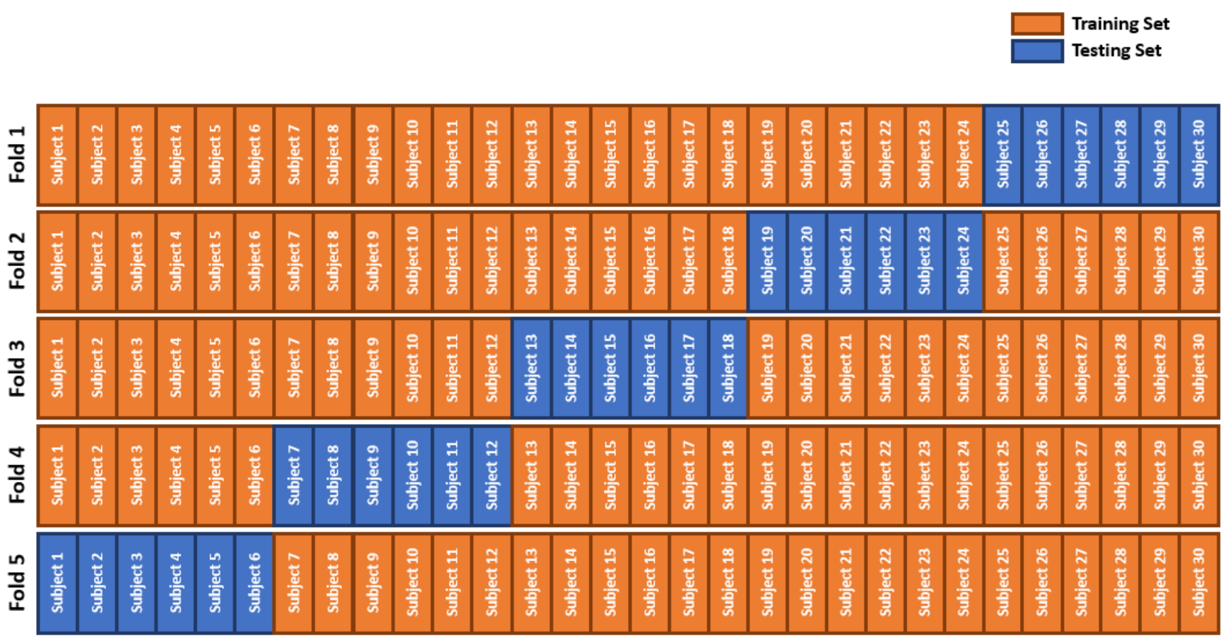


Fig. 38. 5-Fold Cross Validation Training and Testing Sets.

3.8.1 Neural Network (NN)

A sequential neural network with 5 hidden layers was created using the Python Keras machine learning library [105]. Table 15 below describes the design of each sequential layer in the neural network.

TABLE 15. NEURAL NETWORK MODEL ARCHITECTURE

Layer (type)	**Input Shape**	**Output Shape**	**Activation Function**
Dense – Input Layer	(1, #Features)	(1, 64)	ReLu
20% Dropout	(1, 64)	(1, 52)	N/A
Dense	(1, 52)	(1, 64)	ReLu
Dense	(1, 64)	(1, 32)	ReLu
20% Dropout	(1, 32)	(1, 26)	N/A
Dense – Output Layer	(1, 26)	(1, #Classes)	Softmax

Two dropout layers with a 20% dropout rate are used within the model as a regularization method to reduce model overfitting. When compiling the model, the categorical cross-entropy loss function is used since the output is multi-class for the five emotion classes. The Adam optimization function is used to train the model, and the overall accuracy and categorical accuracy metrics are produced to evaluate the effectiveness of the model. Table 16 shows the training parameters for the model as well.

TABLE 16. NEURAL NETWORK TRAINING PARAMETERS

Training Parameter	Values
LOSS FUNCTION	CATEGORICAL CROSS-ENTROPY
OPTIMIZER	ADAM
NUMBER OF EPOCHS	100
BATCH SIZE	10

For each fold of the cross-validation, the model was recreated from scratch and trained for the number of epochs. For five-fold cross-validation, this created five separate models, one for each fold, trained for 100 epochs on a different set of 80% of the subjects and tested on a different set of 20% of the subjects. The models were run on a CUDA-enabled GPU using the Tensorflow Python machine learning backend to save time during training.

3.8.2 Random Forest (RF)

The Random Forest model was hyperparameter tuned with the following values as shown in Table 17. The hyperparameter-tuning and Random Forest model training were performed using the Python SKLearn library [103]. The hyperparameter grid was randomized and different combinations of hyperparameters were tried until a specified number of iterations is reached. For this work, 100 iterations – or combinations of hyperparameters – were tested, and each iteration contained 3-fold cross-validation. The best performing set of hyperparameters from the 100 randomized hyperparameter sets was then saved and is noted in the Results section.

TABLE 17. RANDOM FOREST HYPERPARAMETER TUNING VALUES

Hyperparameter	Values
NUMBER OF ESTIMATORS (TREES)	200, 400, 600, 800, 1000, 1200, 1400, 1600, 1800, 2000
CRITERION	'GINI IMPURITY', 'ENTROPY INFORMATION GAIN'
MAX FEATURES	AUTO, SQUARE ROOT, LOG2
MAX DEPTH	10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 110
MIN SAMPLES SPLIT	2, 5, 10

3.8.3 Support Vector Machine (SVM)

The SVM classifier model was also hyperparameter-tuned to determine the best model for detecting discrete emotions using the CASE dataset, and the hyperparameter grid is shown in Table 18. The hyperparameter-tuning and SVM model training was performed using the Python scikit-learn library

[103]. The hyperparameter grid was randomized and different combinations of hyperparameters were tried until a specified number of iterations is reached. For this work, 100 randomized iterations – or combinations of hyperparameters – were tested, and each iteration contained 3-fold cross-validation. The best performing set of hyperparameters from the 100 randomized hyperparameter sets was then noted and is given in the Results section.

TABLE 18. RANDOM FOREST HYPERPARAMETER TUNING VALUES

Hyperparameter	Values
KERNELS	LINEAR, RBF, POLYNOMIAL
GAMMAS	0.1, 1, 10, 100
C PARAMETER	0.1, 1, 10, 100, 1000
DEGREES	0, 1, 2, 3, 4, 5, 6
COEFFICIENT 0	0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10

3.8.4 1D Convolutional Neural Network (1D-CNN)

The CNN model in this work was constructed with the Python Keras library [105]. Table 19 below shows each layer with the corresponding layer type, input size, output size, and number of parameters for the data. The model was designed based on the Alexnet model converted into a 1D CNN instead of the original 2D CNN [106].

TABLE 19. 1D-CNN MODEL ARCHITECTURE

Layer (type)	Input Shape	Output Shape	Param #
Conv1D	(#Timesteps, #Features)	(10, 96)	18048
BatchNormalization	(10, 96)	(10, 96)	384
Activation – ReLu	(10, 96)	(10, 96)	0
MaxPooling1D	(10, 96)	(5, 96)	0
Conv1D	(5, 96)	(5, 256)	123136
BatchNormalization	(5, 256)	(5, 256)	1024

TABLE 19. CONTINUED

Layer (type)	Input Shape	Output Shape	Param #
Activation – ReLu	(5, 256)	(5, 256)	0
MaxPooling1D	(5, 256)	(2, 256)	0
ZeroPadding1D	(2, 256)	(4, 256)	0
Conv1D	(4, 256)	(4, 512)	393728
BatchNormalization	(4, 512)	(4, 512)	2048
Activation – ReLu	(4, 512)	(4, 512)	0
MaxPooling1D	(4, 512)	(2, 512)	0
ZeroPadding1D	(2, 512)	(4, 512)	0
Conv1D	(4, 512)	(4, 1024)	1573888
BatchNormalization	(4, 1024)	(4, 1024)	4096
Activation – ReLu	(4, 1024)	(4, 1024)	0
ZeroPadding1D	(4, 1024)	(6, 1024)	0
Conv1D	(6, 1024)	(6, 1024)	3146752
BatchNormalization	(6, 1024)	(6, 1024)	4096
Activation	(6, 1024)	(6, 1024)	0
MaxPooling1D	(6, 1024)	(3, 1024)	0
Flatten	(3, 1024)	(3072)	0
Dense	(3072)	(3072)	9440256
BatchNormalization	(3072)	(3072)	12288
Activation	(3072)	(3072)	0
Dropout	(3072)	(3072)	0
Dense	(3072)	(4096)	12587008
BatchNormalization	(4096)	(4096)	16384
Activation	(4096)	(4096)	0
Dropout	(4096)	(4096)	0
Dense	(4096)	(5)	20485
BatchNormalization	(5)	(5)	20
Activation	(5)	(5)	0

The input layer of this model uses L2 regularization on the layer kernel to calculate a least sum squares penalty on the loss function used to optimize this layer. Two 50% dropout layers are also used towards the end of the model to help reduce overfitting. Table 20 shows the training parameters for the model.

TABLE 20. "1D ALEXNET" CNN TRAINING PARAMETERS

Training Parameter	Values
LOSS FUNCTION	CATEGORICAL CROSS-ENTROPY
OPTIMIZER	ADAM
NUMBER OF EPOCHS	100
BATCH SIZE	10

For each fold of the cross-validation, the model was recreated from scratch and trained for the specified number of epochs. For five-fold cross-validation, this created five separate models – one for each fold – trained for 100 epochs on a different set of 80% of the subjects and tested on a different set of 20% of the subjects. The models were run on a CUDA-enabled GPU using the Tensorflow Python machine learning backend to save time during training.

3.8.5 Long Short-Term Memory (LSTM)

The LSTM model was constructed with the Python Keras library [105]. Table 21 below shows each layer with the corresponding layer type, input size, output size, and number of parameters for each layer.

TABLE 21. LSTM MODEL ARCHITECTURE

Layer (type)	**Input Shape**	**Output Shape**	**Param #**
LSTM	(#Timesteps, #Features)	(10, 256)	280576
LSTM	(10, 96)	(256)	525312
Dense	(10, 96)	(256)	65792
Dropout	(10, 96)	(256)	0
Dense	(5, 96)	(64)	16448
Dropout	(5, 256)	(64)	0
Dense	(5, 256)	(64)	4160
Dense	(5, 256)	(32)	2080
Dropout	(2, 256)	(32)	0
Dense	(4, 256)	(32)	1056
Dense	(4, 512)	(5)	165

Two 20% and one 30% dropout layers were used throughout the model to help reduce overfitting. Table 22 shows the training hyperparameters for the model.

TABLE 22. LSTM TRAINING HYPERPARAMETERS

Training Parameter	Values
LOSS FUNCTION	CATEGORICAL CROSS-ENTROPY
OPTIMIZER	ADAM
NUMBER OF EPOCHS	100
BATCH SIZE	50

For each fold of the cross-validation, the model was recreated from scratch and trained for the number of epochs. For five-fold cross-validation, this created five separate models trained for 100 epochs on a different set of 80% of the subjects and tested on a different set of 20% of the subjects. The batch size for this model was increased from 10 to 50 when compared to the NN and CNN models described earlier to reduce the training time of the model since LSTM layers are significantly more computationally expensive than a normal neural network. The models were run on a CUDA-enabled GPU using the Tensorflow Python machine learning backend to save time during training.

CHAPTER 4

RESULTS

4.1 Regression Errors

Table 23 below shows the error results of each regression model. The table is broken up into sections representing the different steps of arousal and valence feature manipulation as described in the Labels Preprocessing section. The lower the error values of the regression model, the higher the performance of the model. In the table below, the error values are color-coded from red (high) to green (low) errors. As can be seen in Table 23, each step in the post-processing of the arousal and valence labels improved the performance of the regression models. The last section in the table: "Scaled Gaussian Oversample" was used in the final method to generate regression models used with the classification model to create the overall ensemble model.

TABLE 23. COLOR-CODED REGRESSION ERROR RESULTS

	Regressor	Valence			Arousal		
		MAE	MSE	RMSE	MAE	MSE	RMSE
Scaled	Linear	0.134	0.033	0.181	0.123	0.027	0.163
	Random Forest	0.154	0.042	0.205	0.138	0.032	0.180
	SVM	0.149	0.040	0.201	0.134	0.031	0.177
	Adaboost	0.134	0.034	0.183	0.123	0.026	0.162
	XG Boost	0.164	0.047	0.218	0.147	0.036	0.190
Scaled Gaussian	Linear	0.140	0.033	0.181	0.128	0.027	0.165
	Random Forest	0.152	0.037	0.193	0.132	0.028	0.167
	SVM	0.149	0.037	0.192	0.135	0.030	0.173
	Adaboost	0.142	0.033	0.182	0.131	0.027	0.166
	XG Boost	0.150	0.037	0.193	0.140	0.032	0.178
Scaled Gaussian Oversample	Linear	0.133	0.030	0.173	0.124	0.026	0.160
	Random Forest	0.130	0.028	0.168	0.119	0.024	0.154
	SVM	0.132	0.030	0.173	0.124	0.026	0.161
	Adaboost	0.138	0.031	0.175	0.126	0.026	0.160
	XG Boost	0.131	0.029	0.171	0.121	0.024	0.156

4.2 Classification Accuracy

Using the methodology explained above, the following graphs give results of multiple different classification models run with arousal and valence features generated from the various regression models above. The features used in each model were the predicted arousal and valence from the respective regressor concatenated with the original physiological features. The ground truth arousal and valence labels concatenated with the physiological feature set as well as just the physiological feature set without arousal and valence features are used as baselines to compare results to the predicted arousal and valence feature sets. The highest accuracy ensemble models among the types and combinations tried were the hyperparameter-tuned SVM classification with linear regression predicted arousal and valence ensemble model with an accuracy of ***98.79% ± 0.29%*** and the neural network with linear regression predicted arousal and valence ensemble model with an accuracy of ***98.33% ± 0.89%***. There were multiple other models with similarly high accuracies of low to mid-90s as shown in Fig. 39 below as well. Multiple metrics are shown to give a more complete representation of the model performance including accuracy, AUC, F1 score, recall, and precision in Fig. 39, Fig. 40, Fig. 41, Fig. 42, and Fig. 43 below respectively. In the metrics, confusion matrix, histogram, learning curve, and other figures below, “No A/V (Only Physiological)” stands for only using the physiological feature set with no arousal or valence features, “True A/V + Phys” stands for the ground truth self-reported arousal and valence labels from the CASE dataset concatenated to the physiological feature set, “Linear Regression A/V + Phys” stands for the predicted arousal and valence values from the linear regressor concatenated to the physiological feature set, and following this convention the others correspond to each regressors’ predicted arousal and valence features concatenated with the physiological feature set respectively.

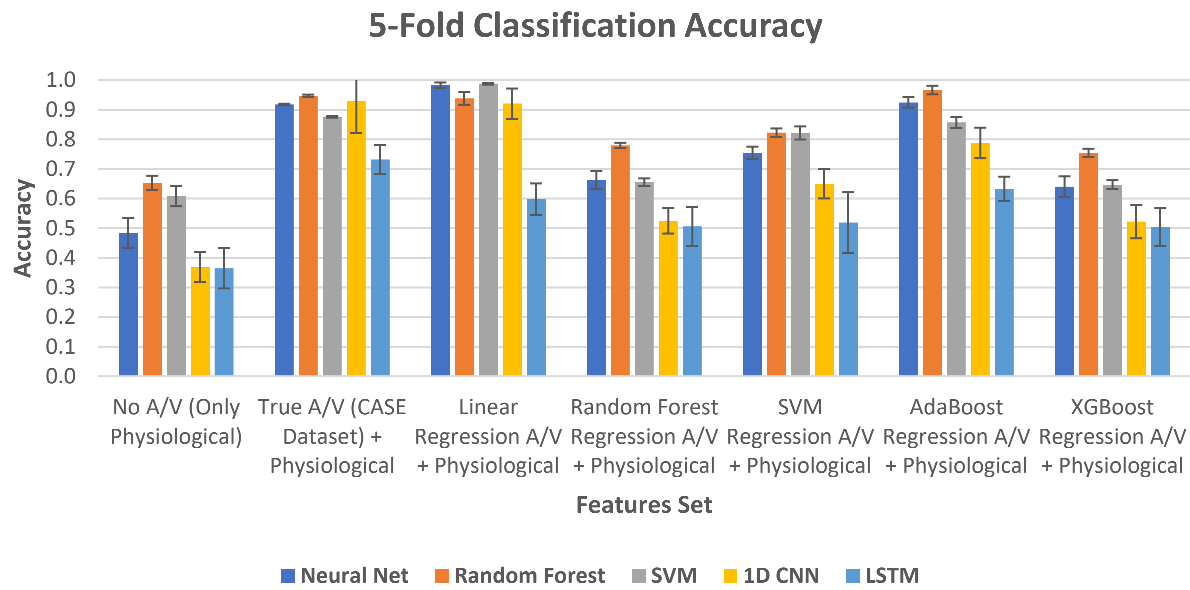


Fig. 39. Classification Model Accuracies w/ each Regressor Predicted Arousal/Valence.

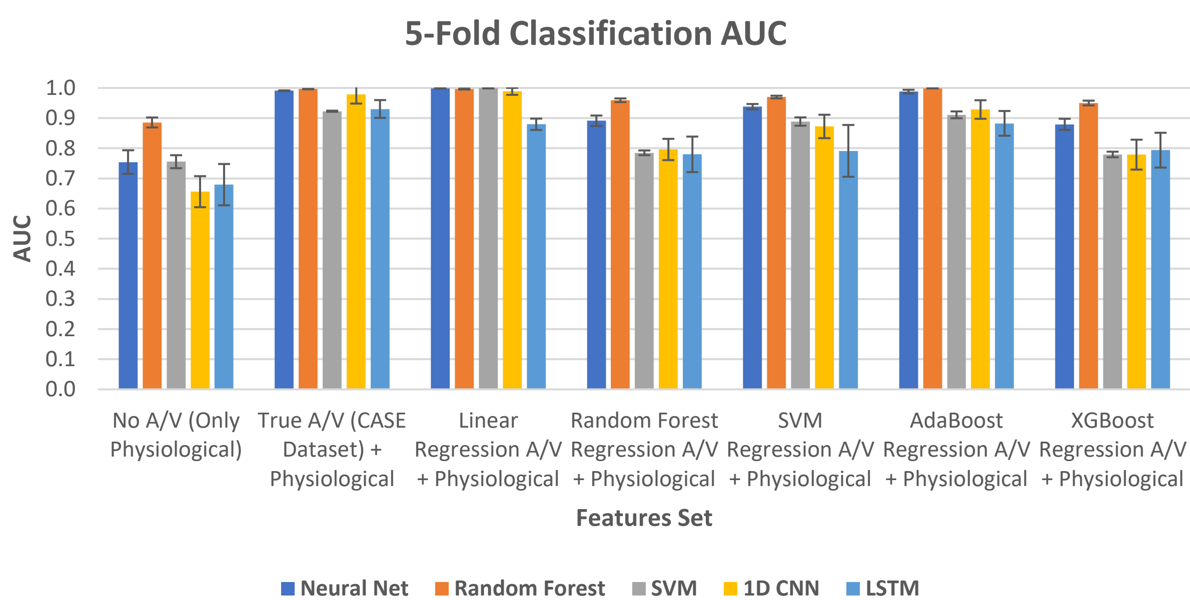


Fig. 40. Classification Model AUCs w/ each Regressor Predicted Arousal/Valence.

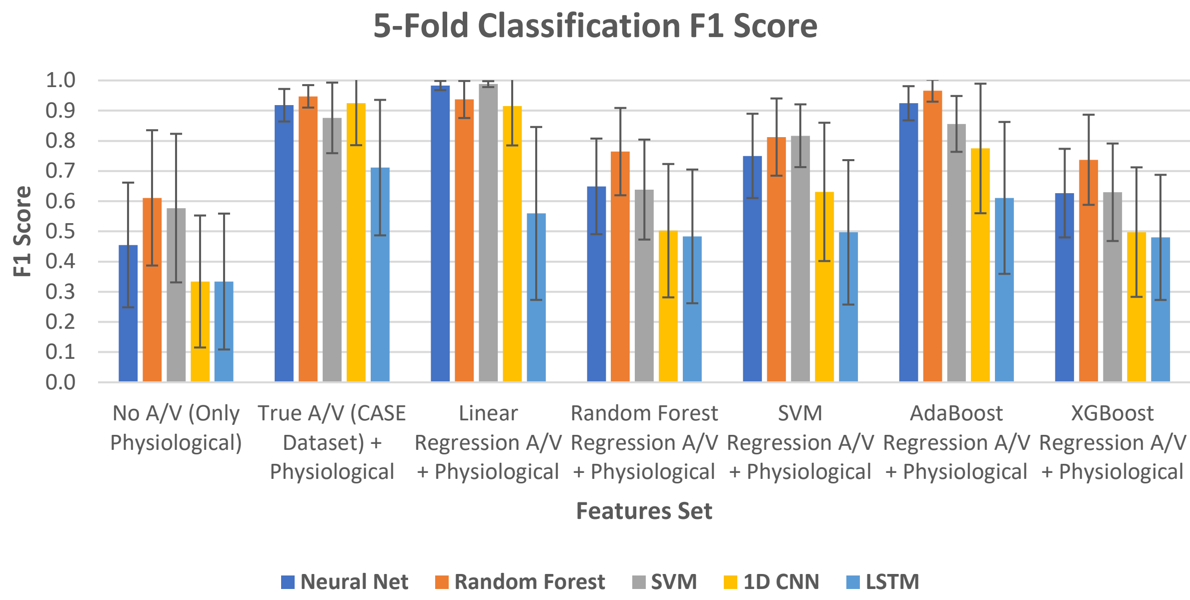


Fig. 41. Classification Model F1 Scores w/ each Regressor Predicted Arousal/Valence.

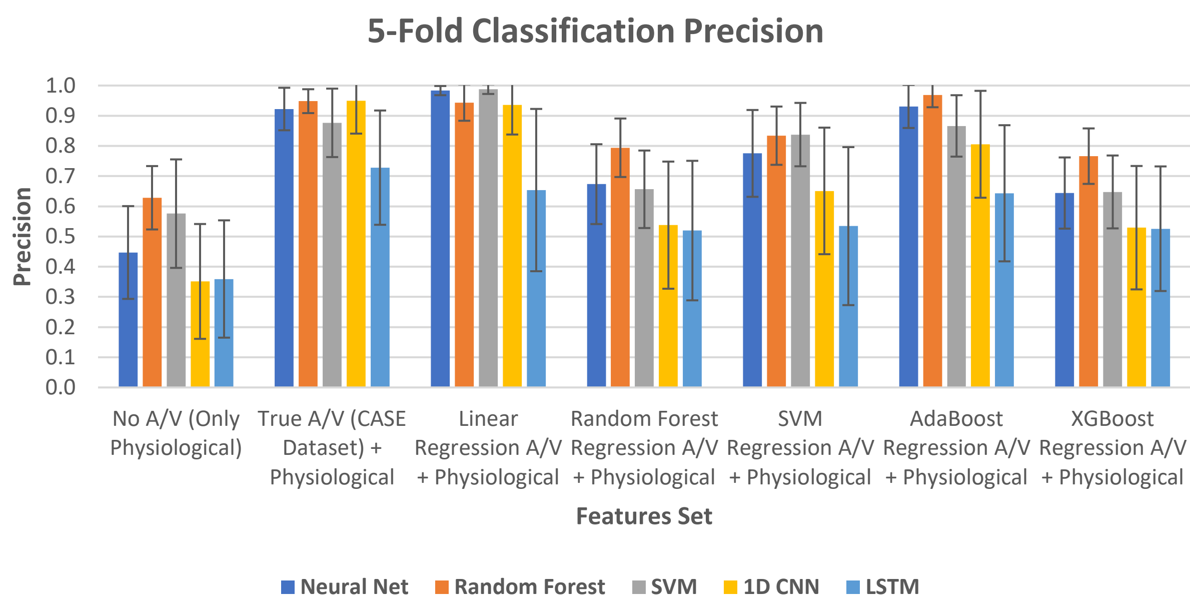


Fig. 42. Classification Model Precisions w/ each Regressor Predicted Arousal/Valence.

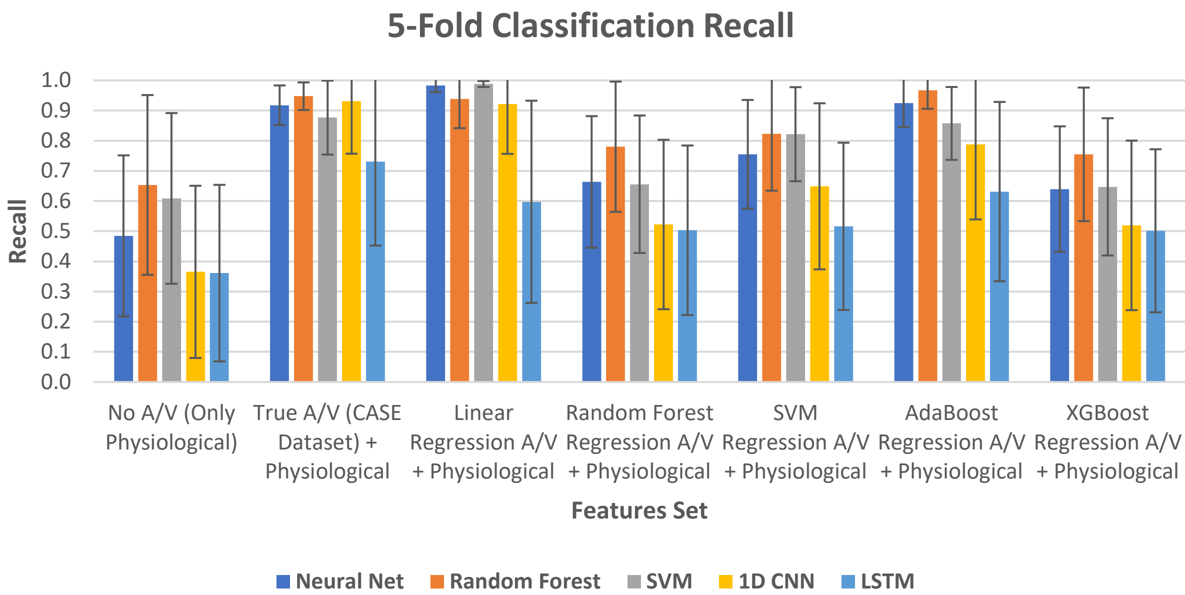


Fig. 43. Classification Model Recalls w/ each Regressor Predicted Arousal/Valence.

The best performing model, SVM with linear regressor arousal and valence prediction, was hyper-parameter-tuned and Table 24 below shows the hyperparameters found to provide the best accuracy among the 100 randomly tried hyperparameter sets from the list of values in Table 18. The decision boundaries of the multiclass SVM as defined by the linear kernel is also shown below in Fig. 44. The SVM model used all 19 features for classification with a 19-dimensional hyperplane, but since it is difficult to visualize 19 dimensions and the two predicted arousal and valence features were the most information-rich, these features were chosen to help visualize the decision boundaries of the models in Fig. 44.

TABLE 24. SVM HYPERPARAMETER-TUNED VALUES FOR SVM WITH LINEAR REGRESSORS ENSEMBLE MODEL

Hyperparameter	Values
KERNELS	LINEAR
GAMMAS	1
C PARAMETER	10
DEGREES	5
COEFFICIENT 0	0

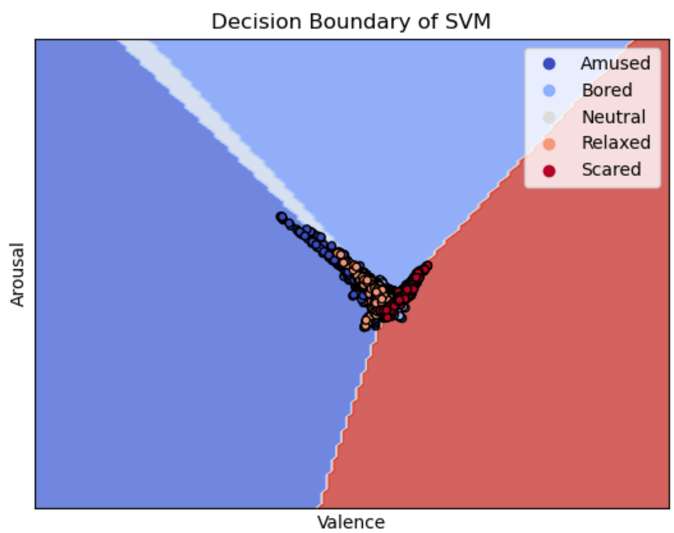


Fig. 44. SVM Decision Boundary with Linear Regressor Predicted Arousal and Valence.

Fig. 45 below shows a comparison of results for the SVM classifier using different feature sets: predicted arousal and valence only (blue bars), physiological features only (grey bar), and predicted arousal and valence concatenated with physiological features (orange bars). As can be seen in Fig. 45, combining the predicted arousal and valence with the physiological features always produced a higher accuracy than just the predicted arousal and valence features by themselves. The grey bar showing the results of only using physiological features gives a baseline for comparison.

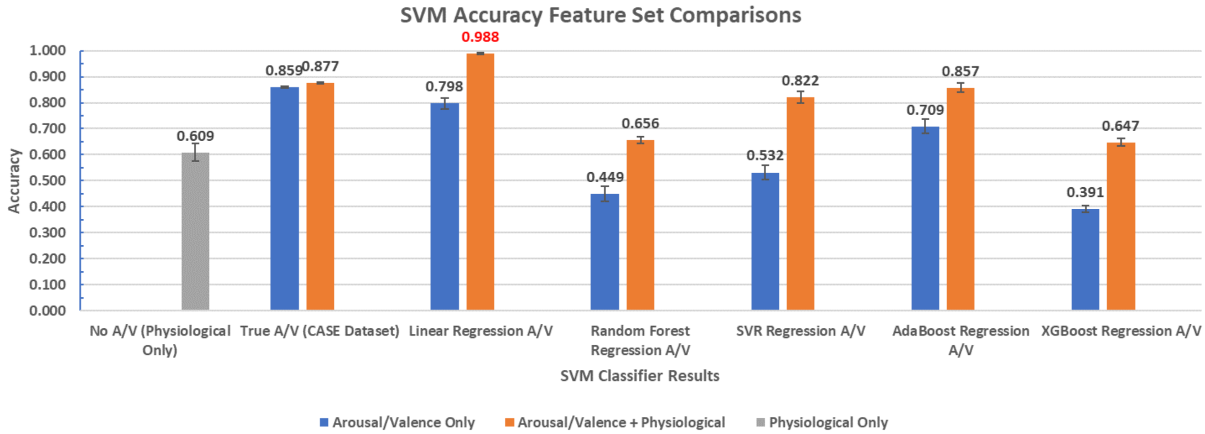


Fig. 45. Comparison of SVM Accuracy Across Different Feature Sets.

4.2.1 Confusion Matrices

Confusion matrices for each classification model trained and tested on each feature set are given below. The confusion matrices show the likelihood of misclassifications between each true and predicted class. The diagonal of the confusion matrix represents correct classification, so the darker the diagonal of a confusion matrix, the better performing the model is. Fig. 46, Fig. 47, Fig. 48, Fig. 49, and Fig. 50 show the confusion matrices for the neural network, random forest, SVM, 1D CNN, and LSTM models respectively for each regression model used to generate predicted arousal and valence features as well as the true arousal and valence from the CASE dataset and no arousal and valence (physiological features only) for baseline comparisons.

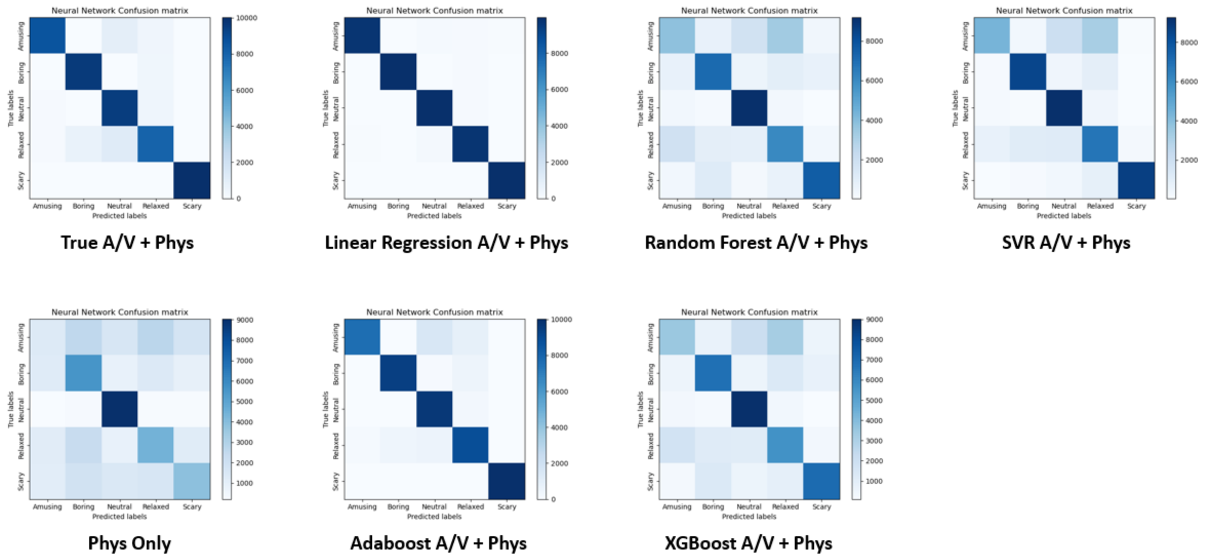


Fig. 46. Neural Network Confusion Matrices.

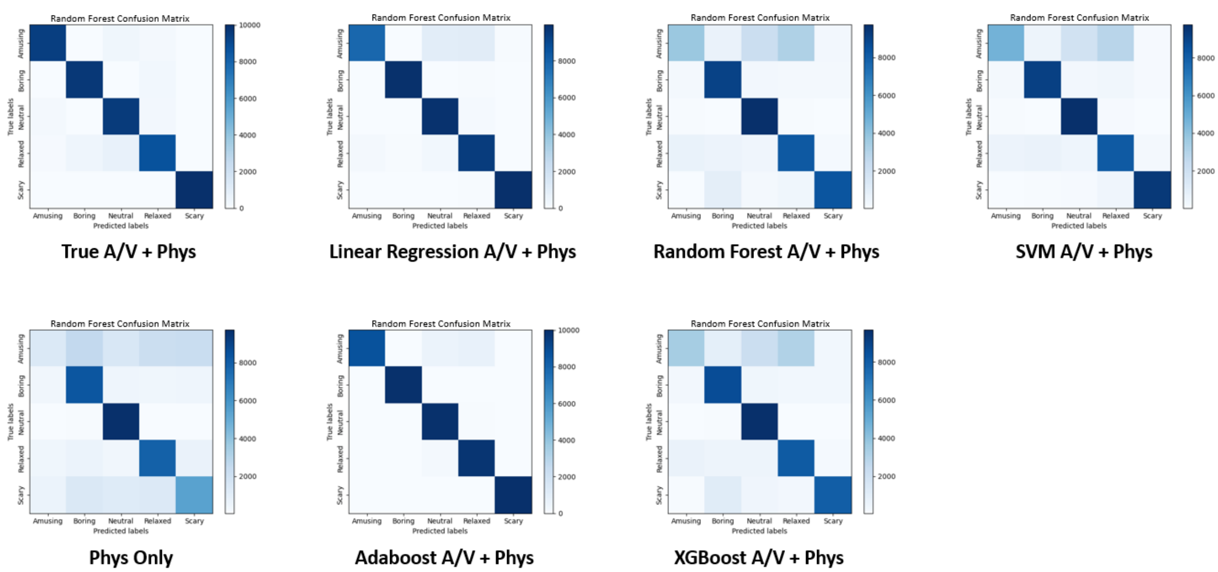


Fig. 47. Random Forest Confusion Matrices.

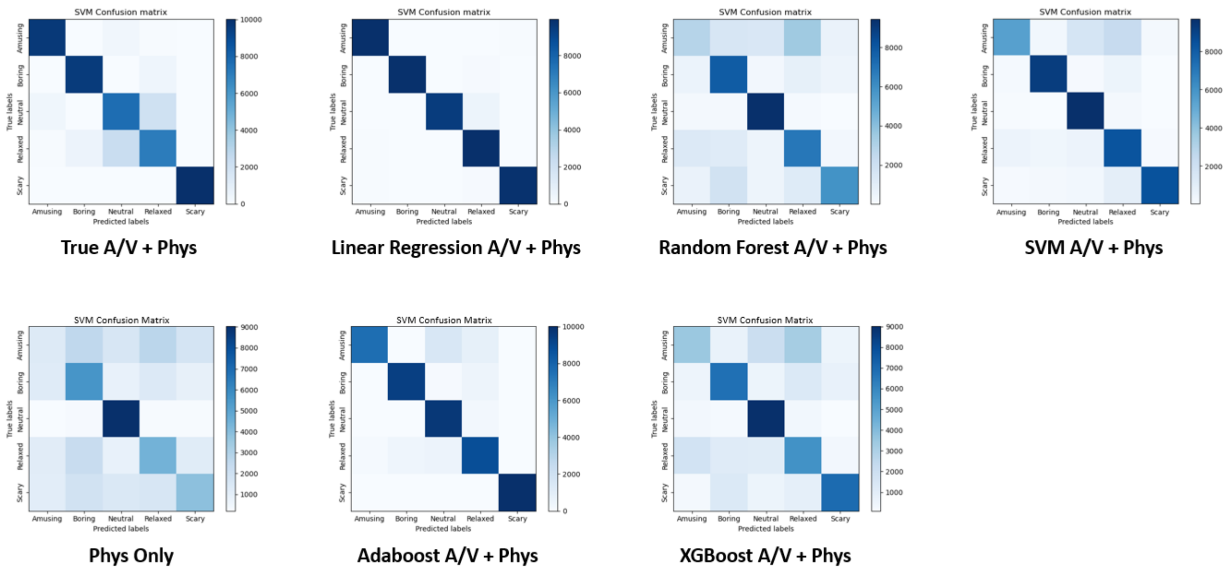


Fig. 48. SVM Confusion Matrices.

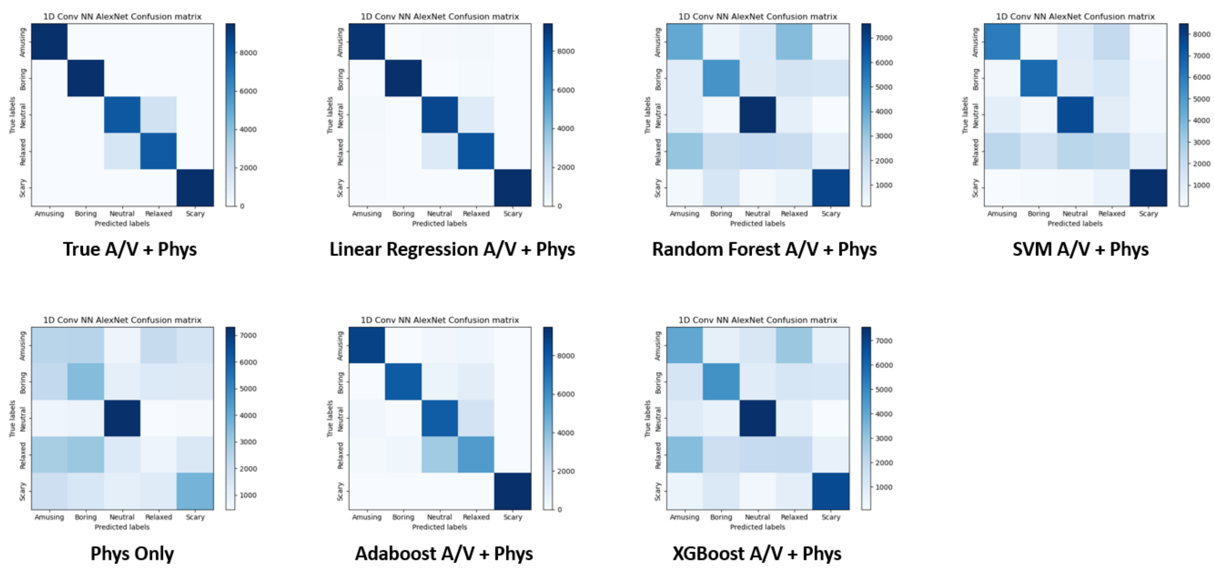


Fig. 49. 1D CNN (1D Alexnet) Confusion Matrices.

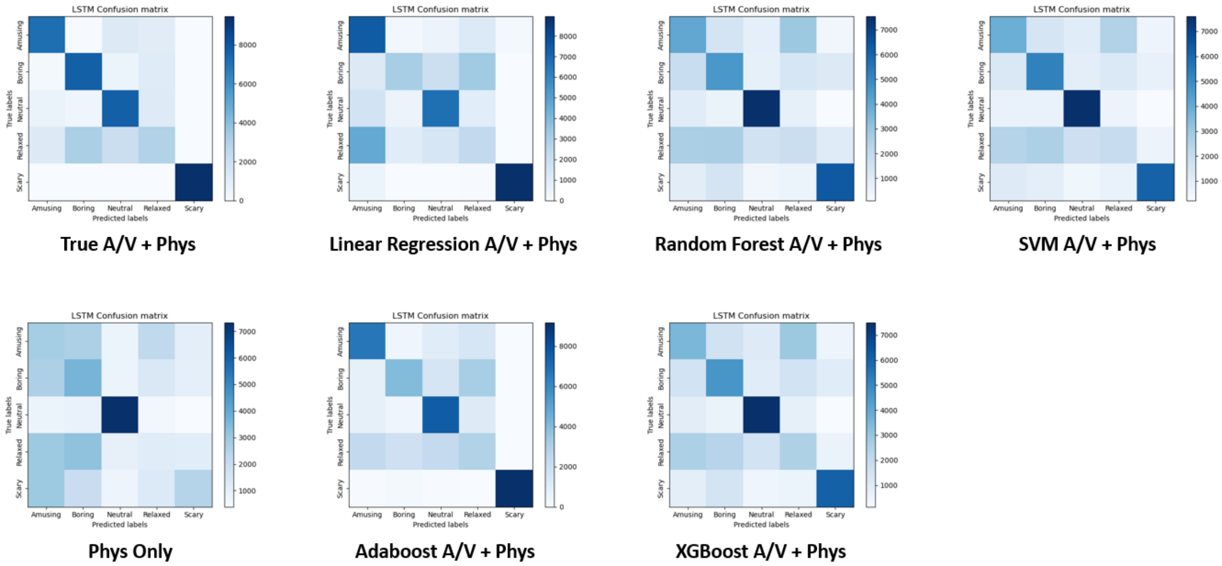


Fig. 50. LSTM Confusion Matrices.

4.2.2 Predicted Class Probability Histograms

For each classification model, the predicted probability values for each prediction of the test dataset were obtained from the corresponding Python library. These class prediction probabilities were then graphed as a histogram for each test dataset in each cross-validation fold to visualize and give insight into the model's "confidence" in its class predictions. In general, the more predictions with a probability greater than 90%, the more accurate the model. Histograms give an idea of the model's confidence in its correct prediction of the class by binning the prediction probability of all windows' features into ten bins and graphing the number of predictions per bin into a histogram. As can be seen in the histograms below, the highest performing ensemble models, neural network with linear regressor and SVM with linear regressor, have almost all predictions in the greater than 90% prediction probability bins. Fig. 51, Fig. 52, Fig. 53, Fig. 54, and Fig. 55 show the histograms of model prediction probability for neural network, random forest, SVM, 1D CNN, and LSTM respectively for each regression model's feature set as well as the self-reported arousal and valence from the CASE dataset and no arousal and valence (physiological features only) for baseline comparisons.

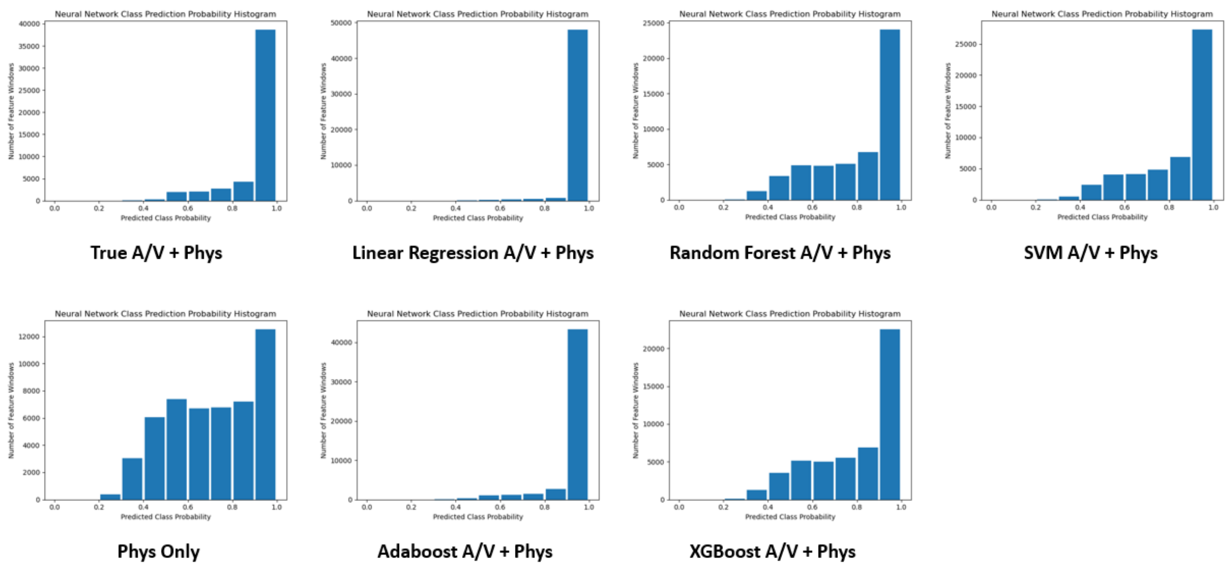


Fig. 51. Neural Network Predicted Class Probability Histogram.

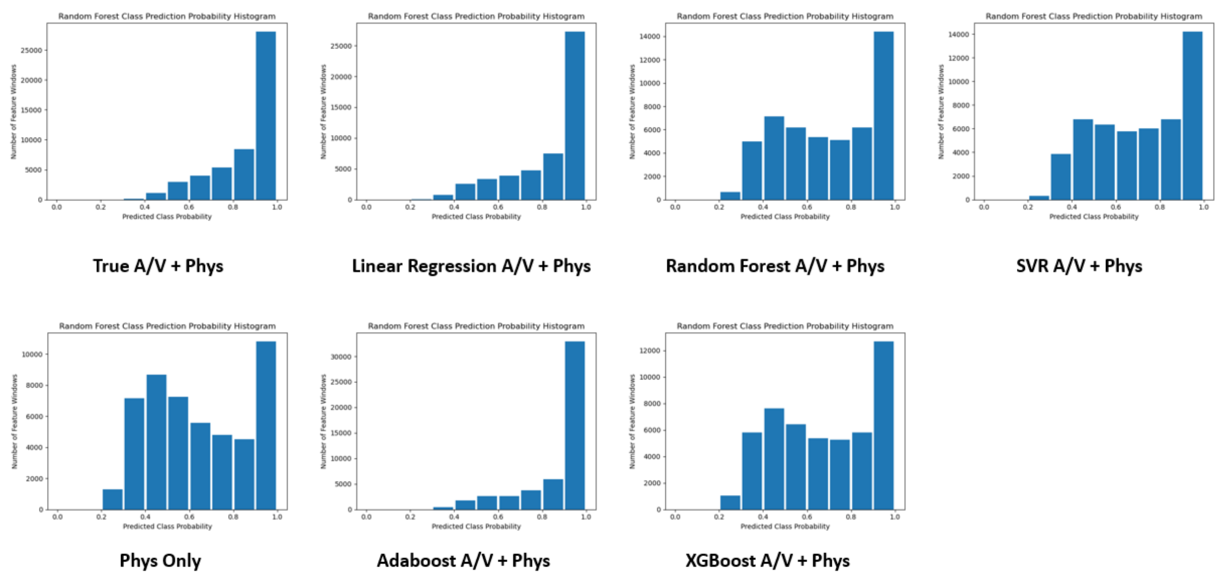


Fig. 52. Random Forest Predicted Class Probability Histogram.

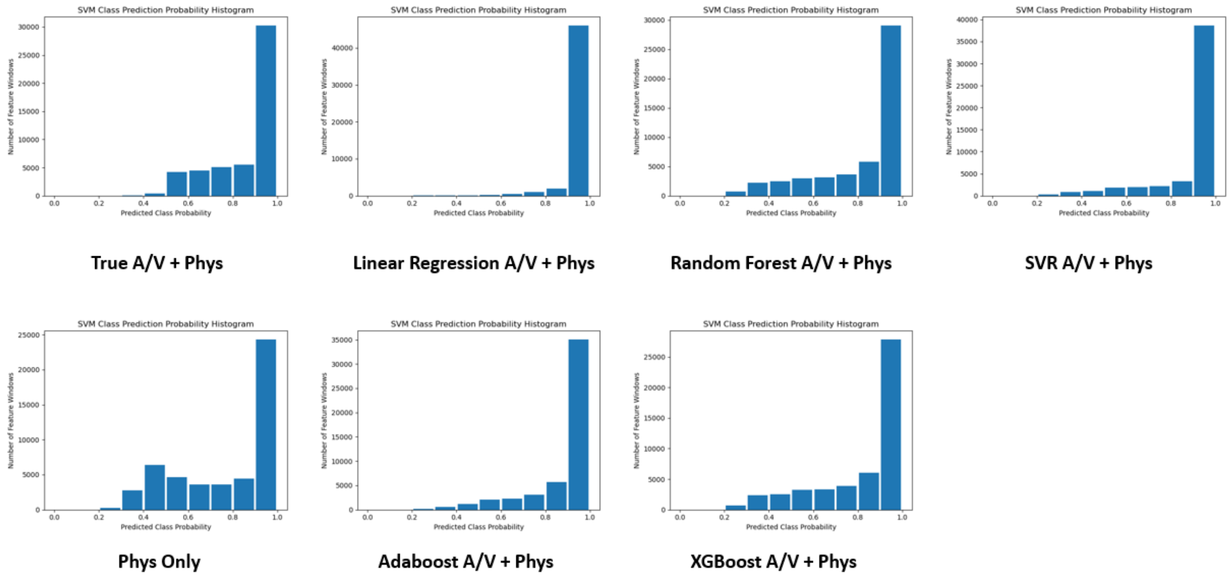


Fig. 53. SVM Predicted Class Probability Histogram.

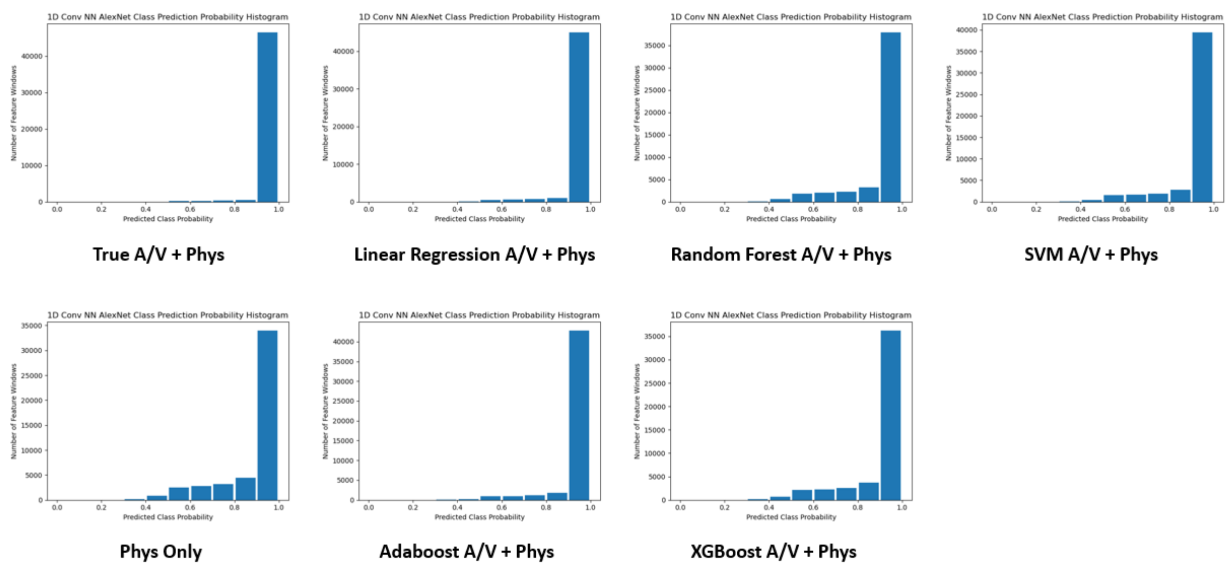


Fig. 54. 1D CNN (1D Alexnet) Predicted Class Probability Histogram.

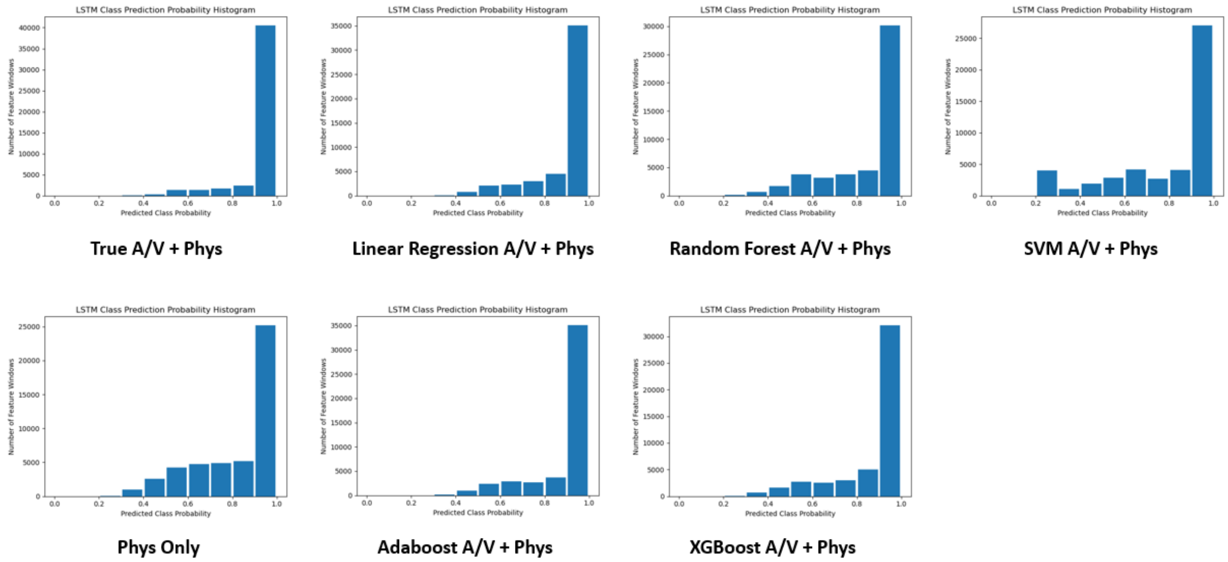


Fig. 55. LSTM Predicted Class Probability Histogram.

4.2.3 Learning Curves

The following graphs for the neural network, 1D CNN, and LSTM models in Fig. 56, Fig. 59, and Fig. 60 show the training accuracy in blue and loss from the corresponding loss function in yellow for each fold in the five-fold cross-validation. The “sixth” fold in the graphs corresponds to a separate model trained on all CASE subjects’ data for use in real-time applications. For the random forest and SVM models in Fig. 57 and Fig. 58, three graphs are given for each model. The top graph shows training and cross-validation accuracies, the middle graph gives an idea of the scalability of the model, and the bottom graph gives an idea of the model’s performance.

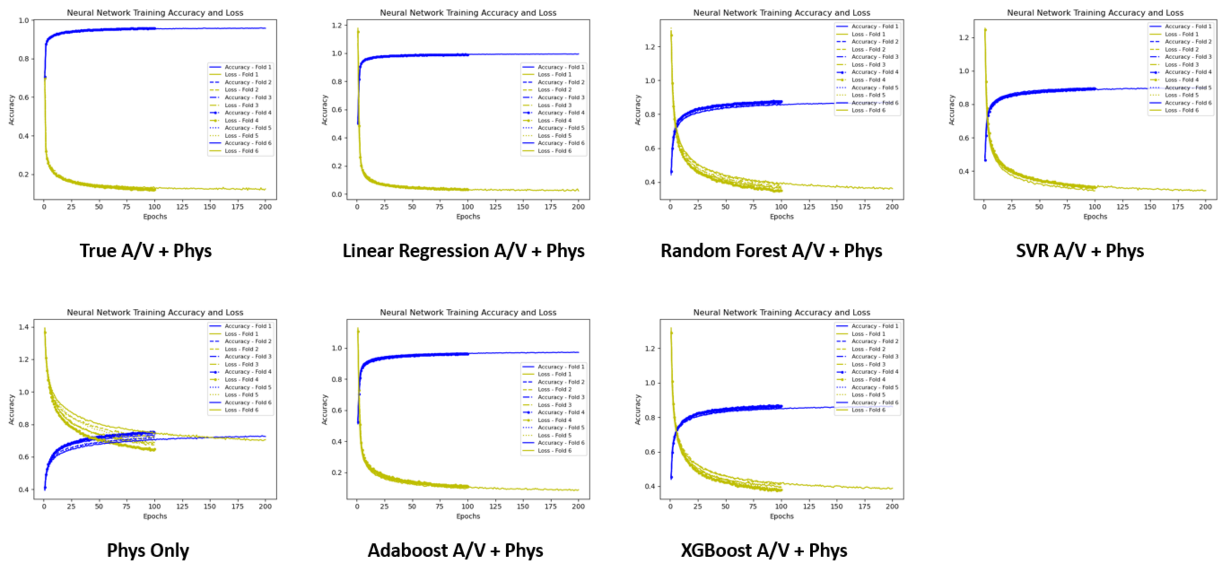
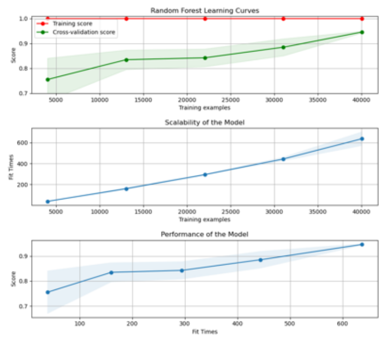
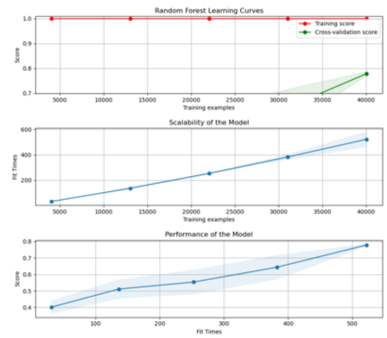


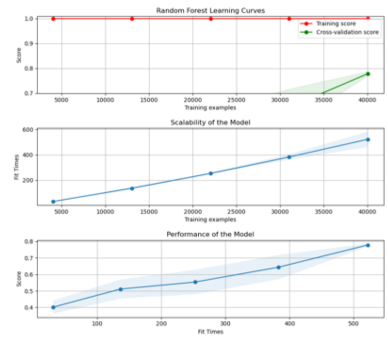
Fig. 56. Neural Network Learning Curves.



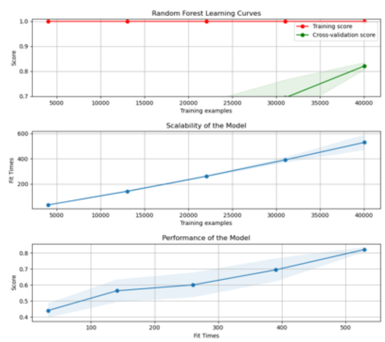
True A/V + Phys



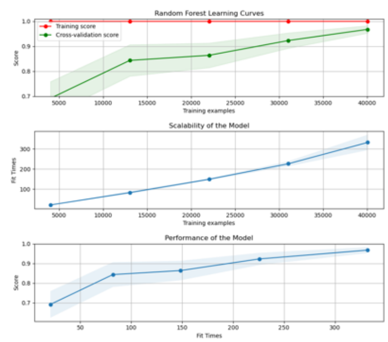
Linear Regression A/V + Phys



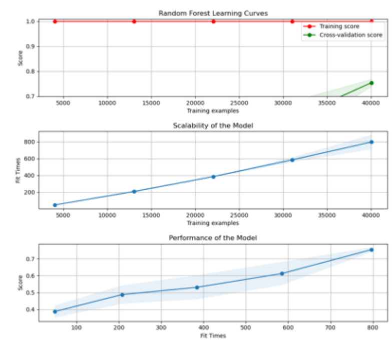
Random Forest A/V + Phys



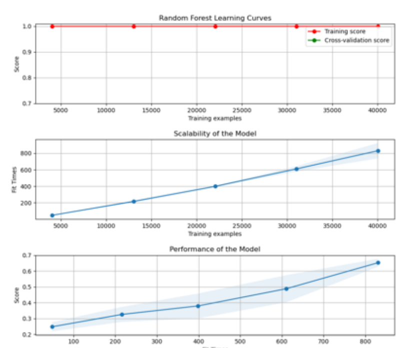
SVR A/V + Phys



Adaboost A/V + Phys

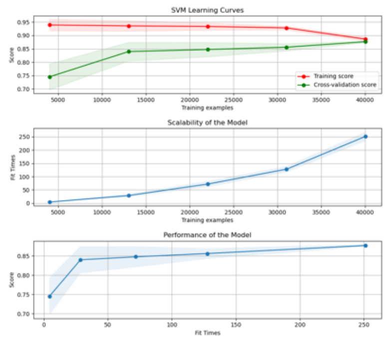


XGBoost A/V + Phys

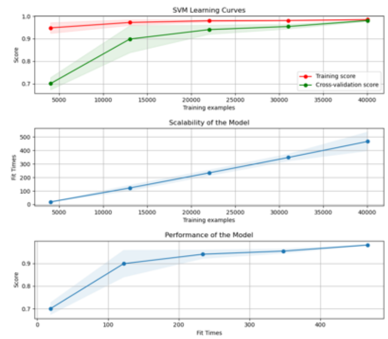


Phys Only

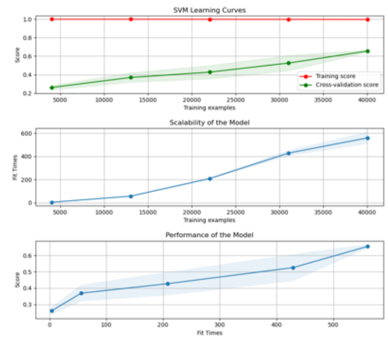
Fig. 57. Random Forest Learning Curves.



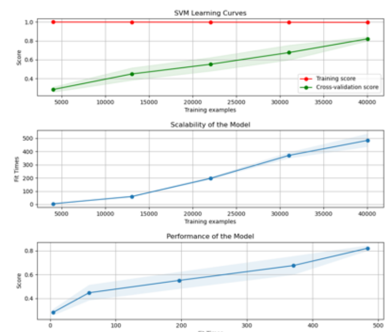
True A/V + Phys



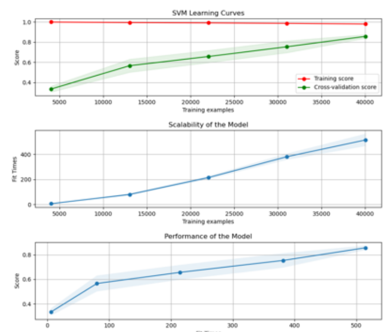
Linear Regression A/V + Phys



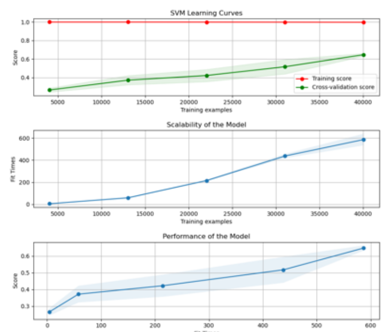
Random Forest A/V + Phys



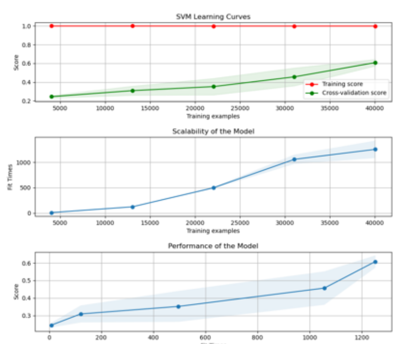
SVR A/V + Phys



Adaboost A/V + Phys



XGBoost A/V + Phys



Phys Only

Fig. 58. SVM Learning Curves.

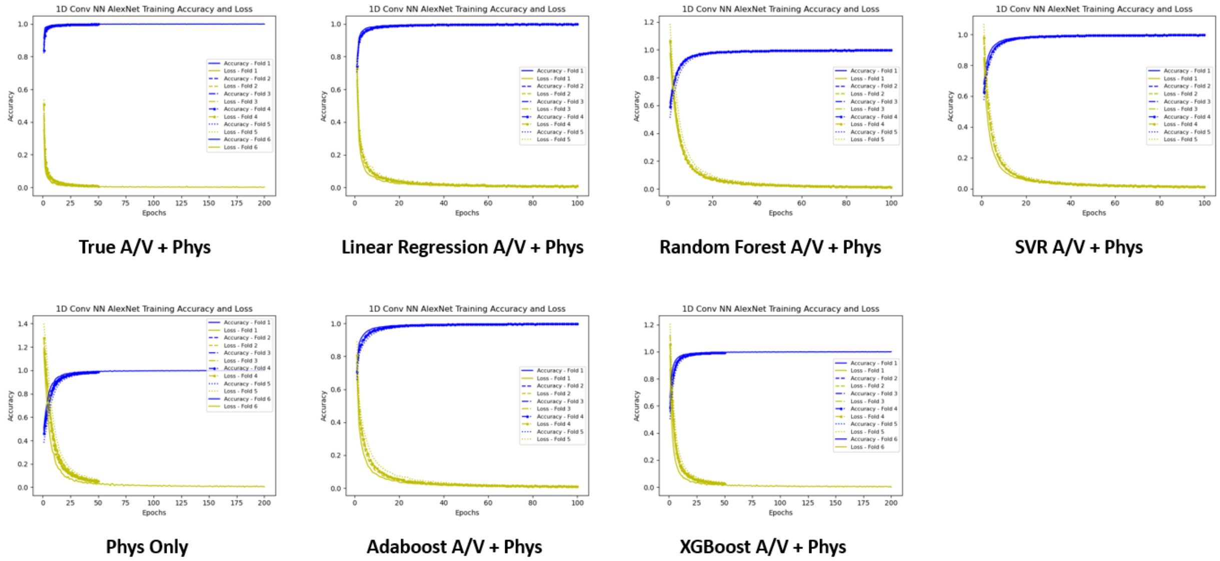


Fig. 59. 1D CNN (1D Alexnet) Learning Curves.

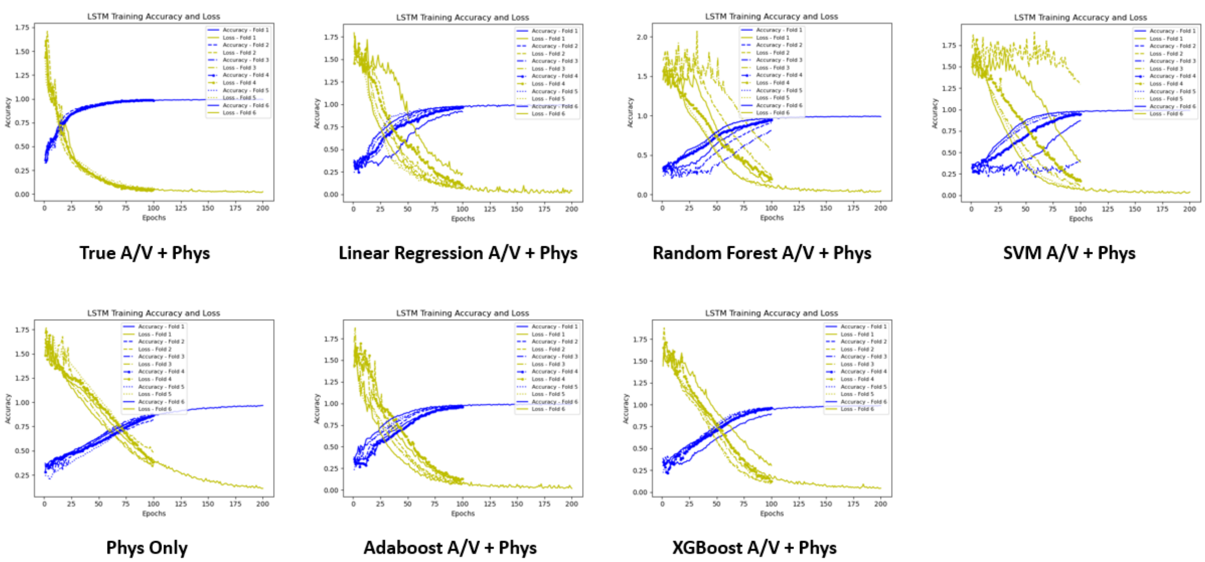


Fig. 60. LSTM Learning Curves.

4.2.4 SVM Decision Boundaries

SVM Decision boundaries along the two dimensions of the arousal and valence features were graphed to provide a visual indication of the model’s decision-making processes in the features’ two-dimensional separation. The SVM model which only used physiological features is not graphed since it did not contain the arousal and valence features. The SVM model used all 19 features for classification with a 19-dimensional hyperplane, but since it is difficult to visualize 19 dimensions and the two predicted arousal and valence features were the most information-rich, these features were chosen to help visualize the decision boundaries of the models. The decision boundaries using these two features generated by each regression model as well as self-annotated arousal and valence for baseline comparison are shown in Fig. 61 below.

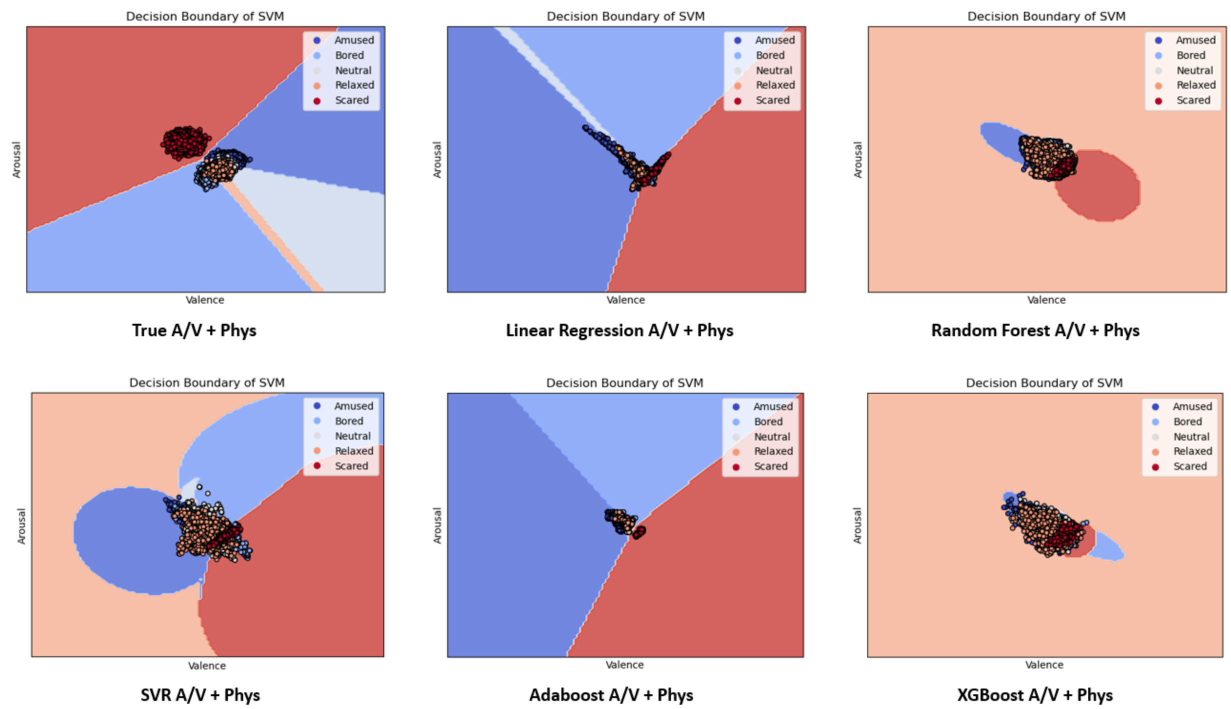


Fig. 61. SVM Decision Boundaries for Models Trained on all Feature Sets.

4.2.5 Preprocessing Method Accuracy Comparisons

To determine the effectiveness of the preprocessing methods employed in this work, ensemble models were trained with various preprocessing steps taken out to compare overall accuracies. These were all trained on the CNN described in the Convolutional Neural Network (CNN) section with a linear regressor to compare just the preprocessing steps and the accuracy results are shown below in Fig. 62. A two-dimensional representation of the CNN's output layer is also shown with the accuracies using t-SNE dimensionality reduction in order to visually see the increase in separation among the five different emotion classes when using the arousal and valence preprocessing methods described in the Labels Preprocessing section.

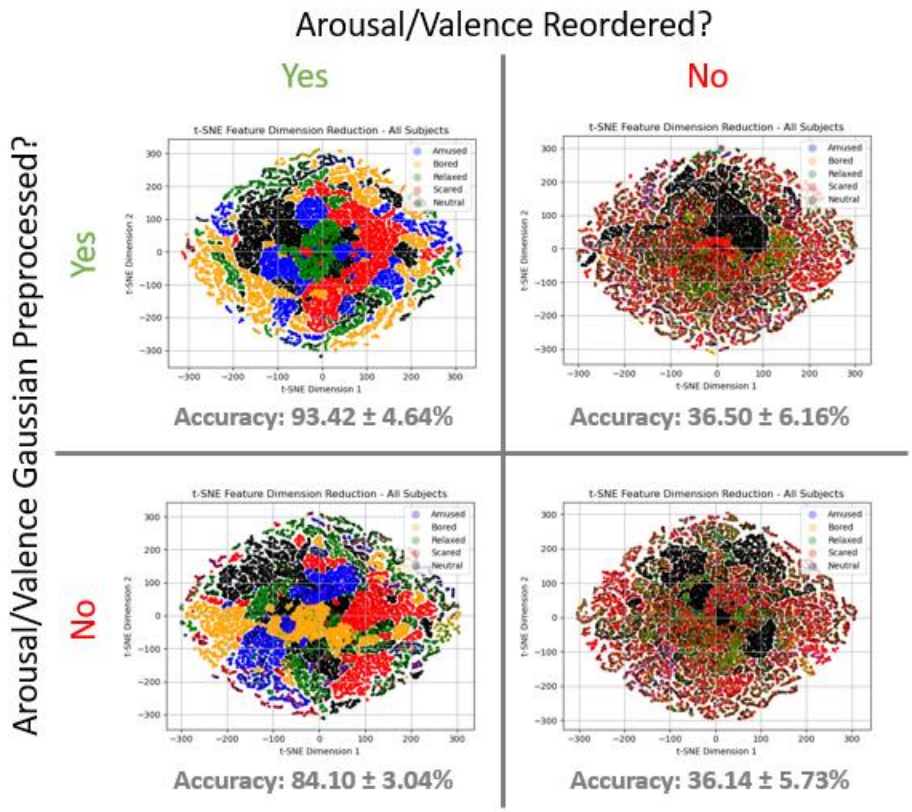


Fig. 62. t-SNE Two-dimensional Representation of CNN Outputs with 5-fold Cross Validation Accuracies to Compare Arousal and Valence Preprocessing Methods.

4.3 Real-time Emotion Detection

The real-time emotion detection pipeline uses the same methodology as the model generation pipeline by windowing the data, filtering the windows, extracting physiological features, predicting arousal and valence using the regressors, and using the combined arousal and valence with the physiological features as inputs to the classifier for predicting discrete emotion labels. The real-time data is acquired from an Empatica E4 physiological sensing device and buffered for 10 seconds with a 1-second window stride. The entire real-time emotion detection workflow using an Empatica E4 device is shown below in Fig. 63.

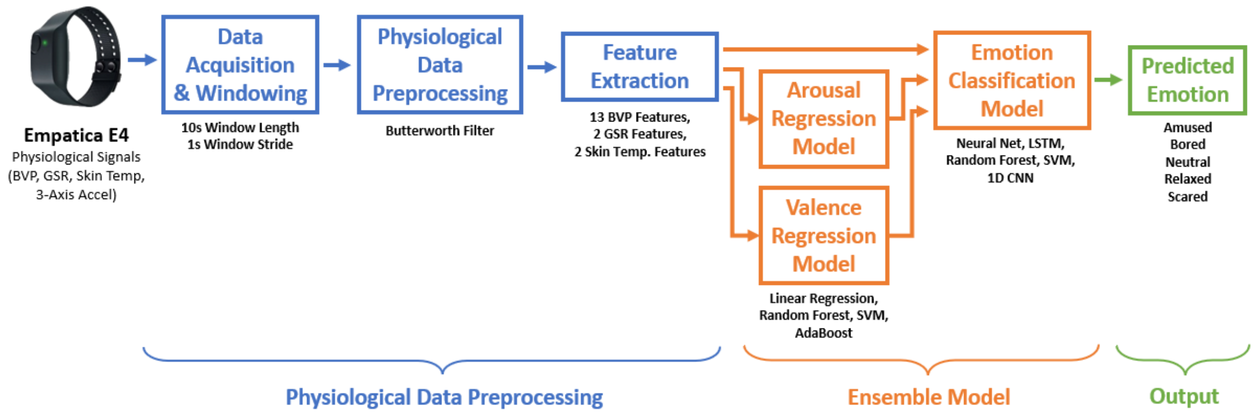


Fig. 63. Real-time Emotion Detection Workflow.

CHAPTER 5

DISCUSSION

A method for detecting five discrete emotions with high accuracy using only signals which are available in affordable, non-invasive, commercial devices has been presented in this work. By using traditional features extracted from physiological data to predict arousal and valence through regression and concatenating the new arousal and valence features with the original physiological features, a much more information-rich feature set can be created. Using the arousal and valence regression models generated in training, only physiological data is needed to create this information-rich feature set in real-world applications of real-time emotion detection. There have been studies found to use regression to create predicted arousal and valence features for emotion classification, but these studies used a different modality (facial expression) to create the arousal and valence features which were then used with physiological features [72] or classified arousal and valence instead of discrete emotion labels [77]. The method developed in this work uses the same physiological data to extract the much more information-rich arousal and valence features than the relatively information-poor physiological features that are traditionally extracted. This is particularly useful because an end-use application would only need the easily attainable physiological data to utilize these models in a real-time environment. A device such as a small smartwatch could easily provide these physiological signals and process the data through the classification methodology with today's technology. Just in this past year, new commercially available devices which detect these three signals have come to the market [107]. It would be fairly simple to develop an app for a smartwatch platform that uses this model as a backend and monitors emotional states in real-time for the user. Similar apps are available today using currently available devices that are used to monitor stress, sleep patterns, and even medical quantities such as heart health using physiological signals from wearable devices.

Table 25 below compares the results from this work to the current state-of-the-art studies in emotion detection.

TABLE 25. ACCURACY COMPARISON WITH CURRENT STATE-OF-THE-ART ALGORITHMS

Reference	Data	Classes	Model	Generalizability	Accuracy
2000, Healey [11]	EMG, BVP, ECG, Respiration, EMG	8-class	K-Nearest Neighbors (KNN)	Subject-dependent (single subject)	81.3%
2005, Herbelin et al. [65]	GSR, BVP, EMG, Respiration, Skin Temperature, Arousal/Valence self-report	5-class	Fisher LDA feature reduction with kNN	Subject-dependent (single subject)	24%
2005, Wagner et al. [64]	EMG, ECG, GSR, Resp	4-class	Linear Discriminant Function (LDF)	Subject-dependent (single subject)	92.1%
2014, Verma and Tiwary [67]	EEG, GSR, BVP, Respiration, Skin Temperature, EMG, EOG	2-class *(Leave one out binary classification of multiple emotions)*	SVM	Subject-independent	77.65% to 85.46% *(Best – Depressing Emotion)*
2019, Albraikan et al. [87]	MAHNOB-HCI	5-class	WMD-DTW and kNN	Subject-independent	65.6%
2020, Liu et al. [71]	GSR	7-class	3-layer Neural Network (NN)	Subject-independent	42.08%
2020, Domínguez-Jiménez [63]	BVP, GSR	3-class	SVM	Subject-dependent	97%
2021, Oh et al. [72]	AffectNet (Facial Expressions) and GSR	8-class	Deep Neural Network (DNN)	Subject-dependent	89%
This Work	***BVP, GSR, Skin Temperature***	***5-class***	***SVM classifier with Linear Regressor***	***Subject-independent***	***98.79%***

To the author’s knowledge, there has not been a method published that uses continuously annotated arousal and valence values to train regression models on physiological features to create predicted arousal and valence features which are then concatenated back to the physiological features to classify discrete emotional states. The achieved accuracies of ***98.79% ± 0.29%*** is also the highest accuracy achieved for subject-independent, five-class discrete emotion classification to the author’s knowledge. Other metrics including AUC, F1 score, precision, and recall were calculated to confirm the validity of the accuracy results of the models, and each metric showed that the models’ performances are valid. A next step would be to test the model on other physiological datasets, but this introduces the challenges of using different emotion models and discrete emotion labels used across datasets as described in the Related Work section. Another next step would be to create more data where the same five emotions are stimulated and test the model using the real-time setup described in the Real-time Emotion Detection section.

There are several notable contributions to the field of affective computing presented in this work. The main contribution is the methodology itself which obtains high accuracies of ~98% with multiple versions of the ensemble model while also being subject-independent and thus generalizable to users outside of the training dataset. Since all testing data for the model contained subjects whose data was never seen during training in each cross-validation fold, the reported accuracy of the model is an accurate indicator of the models' subject independence and generalizability of data acquired outside the CASE dataset. This is a very important distinction since most affective computing research to date has created subject-dependent models which have limited use in real-world applications. As can be seen in the results, the models consistently predicted emotion accurately across all cross-validation folds while being tested on data from subjects that the model had never seen before. The model is also easily implementable in real-world applications since the physiological signals used – BVP, GSR, and skin temperature – can all be acquired through small, non-invasive, commercially available devices. Since the model is a single multi-class model for the five emotion classes, the processing power needed is minimal as well further increasing its utility for use with real-world applications.

A limitation of this model is the demographics of the subjects in the CASE dataset. There is good representation across male and female subjects, but the ages of the subjects are not very dispersed with all subjects being in their 20s and 30s. Also, a potential disadvantage of using discrete emotion labels instead of a continuous emotion space is the phenomenon referred to as emotional "blending" – feeling multiple emotions at the same time [108]. Discrete emotion classification does not adequately represent the complexity of human emotional feelings, so a more nuanced approach may need to be developed in the future for emotion detection problems where the output is able to "blend" emotions instead of placing a single emotional label on a windowed range of time. The continuously predicted arousal and valence features described in this work can be used as more nuanced emotion categorization themselves or used as inputs to more sophisticated "blended" output emotional models. These are ideas for future work presented by the author for the reader to consider.

A potential application of this model is in an environmental emotion feedback system that responds to detected emotions to create a better-suited environment for the occupant. Environments like this have been proposed in other research with various methodologies and use-cases [13-16]. Among other use-cases, an environment like this could help enable differently-abled individuals to live independently and support happier, healthier interactions with their environment. An example interaction could be the system detecting a scared emotional state and outputting soothing music, outputting soothing scents, and dimming the lights. Such a system could help individuals with autism better cope with anxiety associated with bright lights, loud sounds, or other alarming stimuli. An interactive environment such as this could help remove some of the barriers to independent living for differently-abled people.

CHAPTER 6

CONCLUSIONS

This work outlines a method of passive emotion detection using easily acquired, commercially available physiological signals including BVP, GSR, and skin temperature. The model is trained and tested using physiological data, self-reported continuously annotated arousal and valence labels, and emotion labels tied to video stimuli from the CASE dataset [1]. 17 features were extracted from the three physiological signals: thirteen from blood volume pulse (BVP), two from galvanic skin response (GSR), and two from skin temperature. The resulting physiological feature set is used as dependent variables to train two regression models: one for predicting arousal using the self-reported arousal labels, and the other for predicting valence using the self-reported valence labels acquired with the JERI device [96]. The predicted arousal and valence values are then concatenated back to the original physiological feature set to create an aggregated feature set which is used to train and test a classification model for discrete emotion prediction based on the emotion labels from the CASE video stimuli. The regression and classification models together constitute the ensemble model presented in this work. Data from 30 subjects were used for training and testing the ensemble model, and five-fold cross-validation where each fold does not share data from the subjects it contains is used to ensure that the results given are indicative of the accuracy expected of the model from subjects that it has never seen before. The model is a multi-class predictor that classifies the windowed data into five emotion classes: amused, bored, neutral, relaxed, and scared. The best performing model is the SVM classifier using a linear regressor for the arousal and valence prediction features with a five-fold cross-validation accuracy of ***98.79% ± 0.29%***. A real-time implementation of this system is also presented for use with future work utilizing the Empatica E4 device.

Some possible future work with this model could be recording data from more subjects with the same emotion eliciting videos as was used in the CASE dataset with different populations including differently-abled populations. This model can also be used in an emotion feedback environment which could be particularly useful for differently-abled individuals to help enable them to live independent, healthy, and happy lives.

REFERENCES

[1] K. Sharma, C. Castellini, E. L. van den Broek, A. Albu-Schaeffer, and F. Schwenker, "A dataset of continuous affect annotations and physiological signals for emotion analysis," *Scientific data,* vol. 6, no. 1, pp. 1-13, 2019.

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

[3] A. C. M. Bedekar, "Human Emotion Recognition using Physiological Signals: A Survey," *SSRN Electronic Journal*, 2020, doi: 10.2139/ssrn.3645402.

[4] M. Egger, M. Ley, and S. Hanke, "Emotion Recognition from Physiological Signal Analysis: A Review," *Electronic Notes in Theoretical Computer Science,* vol. 343, pp. 35-55, 2019/05/04/ 2019, doi: https://doi.org/10.1016/j.entcs.2019.04.009.

[5] O. Faust, Y. Hagiwara, T. J. Hong, O. S. Lih, and U. R. Acharya, "Deep learning for healthcare applications based on physiological signals: A review," *Computer Methods and Programs in Biomedicine,* vol. 161, pp. 1-13, 2018/07/01/ 2018, doi: https://doi.org/10.1016/j.cmpb.2018.04.005.

[6] S. Greene, H. Thapliyal, and A. Caban-Holt, "A survey of affective computing for stress detection: Evaluating technologies in stress detection for better health," *IEEE Consumer Electronics Magazine,* vol. 5, no. 4, pp. 44-56, 2016.

[7] S. Jerritta, M. Murugappan, R. Nagarajan, and K. Wan, "Physiological signals based human emotion recognition: a review," in *IEEE 7th International Colloquium on Signal Processing and its Applications*, 2011 2011: IEEE, pp. 410-415, doi: 10.1109/CSPA.2011.5759912.

[8] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, "A review of affective computing: From unimodal analysis to multimodal fusion," *Information Fusion,* vol. 37, pp. 98-125, 2017.

[9] L. Shu *et al.*, "A Review of Emotion Recognition Using Physiological Signals," *Sensors,* vol. 18, no. 7, 2018, doi: 10.3390/s18072074.

[10] J. Zhang, Z. Yin, P. Chen, and S. Nichele, "Emotion recognition using multi-modal data and machine learning techniques: A tutorial and review," *Information Fusion,* vol. 59, pp. 103-126, 2020/07/01/ 2020, doi: https://doi.org/10.1016/j.inffus.2020.01.011.

[11] J. A. Healey, "Wearable and automotive systems for affect recognition from physiology," PhD Dissertation, Electrical Engineering and Computer Science, Massachusetts Institute of Technology, 2000.

[12] M. M. Seltzer, M. W. Krauss, P. T. Shattuck, G. Orsmond, A. Swe, and C. Lord, "The symptoms of autism spectrum disorders in adolescence and adulthood," *Journal of autism and developmental disorders,* vol. 33, no. 6, pp. 565-581, 2003.

[13] R. W. Picard and J. Healey, "Affective wearables," *Personal technologies,* vol. 1, no. 4, pp. 231-240, 1997.

[14] A. Fernández-Caballero *et al.*, "Smart environment architecture for emotion detection and regulation," *Journal of biomedical informatics,* vol. 64, pp. 55-73, 2016.

[15] A. Albraikan, B. Hafidh, and A. El Saddik, "iAware: A real-time emotional biofeedback system based on physiological signals," *IEEE Access,* vol. 6, pp. 78780-78789, 2018.

[16] P. Sundaravadivel, V. Goyal, and L. Tamil, "i-rise: An iot-based semi-immersive affective monitoring framework for anxiety disorders," in *EEE International Conference on Consumer Electronics (ICCE)*, 2020 2020: IEEE, pp. 1-5.

[17] M. Gendron and L. Feldman Barrett, "Reconstructing the past: A century of ideas about emotion in psychology," *Emotion review,* vol. 1, no. 4, pp. 316-339, 2009.

[18] J. Panksepp, *Affective neuroscience: The foundations of human and animal emotions*. Oxford university press, 2004.

[19] S. Ps and G. Mahalakshmi, "Emotion models: a review," *International Journal of Control Theory and Applications,* vol. 10, pp. 651-657, 2017.

[20] A. Mehrabian, "Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament," *Current Psychology,* vol. 14, no. 4, pp. 261-292, 1996.

[21] R. Plutchik, "The nature of emotions: Clinical implications," in *Emotions and psychopathology*: Springer, 1988, pp. 1-20.

[22] W. M. Winton, L. E. Putnam, and R. M. Krauss, "Facial and autonomic manifestations of the dimensional structure of emotion," *Journal of Experimental Social Psychology,* vol. 20, no. 3, pp. 195-216, 1984.

[23] S. Koelstra *et al.*, "Deap: A database for emotion analysis; using physiological signals," *IEEE transactions on affective computing,* vol. 3, no. 1, pp. 18-31, 2011.

[24] J. R. J. Fontaine, K. R. Scherer, E. B. Roesch, and P. C. Ellsworth, "The world of emotions is not two-dimensional," *Psychological science,* vol. 18, no. 12, pp. 1050-1057, 2007.

[25] C. M. Lee and S. S. Narayanan, "Toward detecting emotions in spoken dialogs," *IEEE transactions on speech and audio processing,* vol. 13, no. 2, pp. 293-303, 2005.

[26] J.-C. Martin, G. Caridakis, L. Devillers, K. Karpouzis, and S. Abrilian, "Manual annotation and automatic image processing of multimodal emotional behaviors in tv interviews," in *IFIP International Conference on Artificial Intelligence Applications and Innovations*, 2006: Springer, pp. 369-377.

[27] T. Vogt, E. André, and N. Bee, "EmoVoice—A framework for online recognition of emotions from voice," in *International Tutorial and Research Workshop on Perception and Interactive Technologies for Speech-Based Systems*, 2008: Springer, pp. 188-199.

[28] M. Hasan, E. Rundensteiner, and E. Agu, "Emotex: Detecting emotions in twitter messages," 2014.

[29] P. Ekman, "Basic emotions," *Handbook of cognition and emotion,* vol. 98, no. 45-60, p. 16, 1999.

[30] C. E. Izard, "Basic emotions, relations among emotions, and emotion-cognition relations," 1992.

[31] R. W. Levenson, "Human emotion: A functional view," *The nature of emotion: Fundamental questions,* vol. 1, pp. 123-126, 1994.

[32] J. L. Tracy and D. Randles, "Four models of basic emotions: a review of Ekman and Cordaro, Izard, Levenson, and Panksepp and Watt," *Emotion review,* vol. 3, no. 4, pp. 397-405, 2011.

[33] C. O. Alm, D. Roth, and R. Sproat, "Emotions from text: machine learning for text-based emotion prediction," in *Proceedings of human language technology conference and conference on empirical methods in natural language processing*, 2005, pp. 579-586.

[34] C. Strapparava and R. Mihalcea, "Learning to identify emotions in text," in *ACM symposium on Applied computing*, 2008 2008, pp. 1556-1560.

[35] A. J. Gill, R. M. French, D. Gergle, and J. Oberlander, "Identifying emotional characteristics from short blog texts," in *30th Annual Conference of the Cognitive Science Society*, 2008: Citeseer, pp. 2237-2242.

[36] A. Balahur, J. M. Hermida, and A. Montoyo, "Detecting implicit expressions of sentiment in text based on commonsense knowledge," in *2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA 2.011)*, 2011, pp. 53-60.

[37] R. C. Balabantaray, M. Mohammad, and N. Sharma, "Multi-class twitter emotion classification: A new approach," *International Journal of Applied Information Systems,* vol. 4, no. 1, pp. 48-53, 2012.

[38] K. Roberts, M. A. Roach, J. Johnson, J. Guthrie, and S. Harabagiu, "Empatweet: Annotating and detecting emotions on twitter," in *Eighth International Conference on Language Resources and Evaluation (LREC'12)*, 2012, pp. 3806-3813.

[39] A. Agrawal and A. An, "Unsupervised emotion detection from text using semantic and syntactic relations," in *IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology*, 2012, vol. 1: IEEE, pp. 346-353.

[40] M. Sykora, T. Jackson, A. O'Brien, and S. Elayan, "Emotive ontology: Extracting fine-grained emotions from terse, informal messages," 2013.

[41] X. Wang and Q. Zheng, "Text emotion classification research based on improved latent semantic analysis algorithm," in *Proceedings of the 2nd International Conference on Computer Science and Electronics Engineering (ICCSEE 2013)*, 2013: Citeseer, pp. 210-213.

[42] J. Suttles and N. Ide, "Distant supervision for emotion classification with discrete binary values," in *International Conference on Intelligent Text Processing and Computational Linguistics*, 2013: Springer, pp. 121-136.

[43] R. A. Calvo and S. Mac Kim, "Emotions in text: dimensional and categorical models," *Computational Intelligence*, vol. 29, no. 3, pp. 527-543, 2013.

[44] G. McKeown, M. Valstar, R. Cowie, M. Pantic, and M. Schroder, "The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent," *IEEE transactions on affective computing*, vol. 3, no. 1, pp. 5-17, 2011.

[45] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions," in *10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, 2013 2013: IEEE, pp. 1-8.

[46] M. K. Abadi, R. Subramanian, S. M. Kia, P. Avesani, I. Patras, and N. Sebe, "DECAF: MEG-based multimodal database for decoding affective physiological responses," *IEEE Transactions on Affective Computing*, vol. 6, no. 3, pp. 209-222, 2015.

[47] J. Kossaifi *et al.*, "Sewa db: A rich database for audio-visual emotion and sentiment research in the wild," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 3, pp. 1022-1040, 2019.

[48] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic, "A multimodal database for affect recognition and implicit tagging," *IEEE transactions on affective computing*, vol. 3, no. 1, pp. 42-55, 2011.

[49] S. Katsigiannis and N. Ramzan, "DREAMER: A database for emotion recognition through EEG and ECG signals from wireless low-cost off-the-shelf devices," *IEEE journal of biomedical and health informatics*, vol. 22, no. 1, pp. 98-107, 2017.

[50] J. A. Miranda-Correa, M. K. Abadi, N. Sebe, and I. Patras, "Amigos: A dataset for affect, personality and mood research on individuals and groups," *IEEE Transactions on Affective Computing*, vol. 12, no. 2, pp. 479-493, 2018.

[51] R. D. Snee, "Developments in Linear Regression Methodology: 1959-1982," *Technometrics*, vol. 25, no. 3, pp. 230-237, 1983.

[52] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5-32, 2001.

[53] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *Journal of computer and system sciences*, vol. 55, no. 1, pp. 119-139, 1997.

[54] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785-794.

[55] O. D. Team, "Introduction to Support Vector Machines." [Online]. Available: https://docs.opencv.org/2.4/doc/tutorials/ml/introduction_to_svm/introduction_to_svm.html#goal%3E

[56] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The journal of machine learning research*, vol. 15, no. 1, pp. 1929-1958, 2014.

[57] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE transactions on neural networks*, vol. 5, no. 2, pp. 157-166, 1994.

[58] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735-1780, 1997.

[59] J. T. Cacioppo, R. E. Petty, M. E. Losch, and H. S. Kim, "Electromyographic activity over facial muscle regions can differentiate the valence and intensity of affective reactions," *Journal of personality and social psychology*, vol. 50, no. 2, p. 260, 1986.

[60] R. Fernandez, "Stochastic modeling of physiological signals with hidden markov models: A step toward frustration detection in human-computer interfaces," 1998.

[61] J. Healey and R. Picard, "Digital processing of affective signals," 1998 1998, vol. 6: IEEE, pp. 3749-3752.

[62] E. Vyzas and R. W. Picard, "Affective pattern classification," *Emotional and Intelligent: The Tangled Knot of Cognition,* vol. 176182, 1998.

[63] J. A. Domínguez-Jiménez, K. C. Campo-Landines, J. C. Martínez-Santos, E. J. Delahoz, and S. H. Contreras-Ortiz, "A machine learning model for emotion recognition from physiological signals," *Biomedical signal processing and control,* vol. 55, p. 101646, 2020.

[64] J. Wagner, J. Kim, and E. André, "From physiological signals to emotions: Implementing and comparing selected methods for feature extraction and classification," 2005 2005: IEEE, pp. 940-943.

[65] B. Herbelin, P. Benzaki, F. Riquier, O. Renault, H. Grillon, and D. Thalmann, "Using physiological measures for emotional assessment: a computer-aided tool for cognitive and behavioural therapy," *International Journal on Disability and Human Development,* vol. 4, no. 4, pp. 269-278, 2005.

[66] C. Maaoui and A. Pruski, "A comparative study of SVM kernel applied to emotion recognition from physiological signals," in *5th International Multi-Conference on Systems, Signals and Devices*, 2008 2008: IEEE, pp. 1-6.

[67] G. K. Verma and U. S. Tiwary, "Multimodal fusion framework: A multiresolution approach for emotion classification and recognition from physiological signals," *NeuroImage,* vol. 102, pp. 162-172, 2014.

[68] W. Wen, G. Liu, N. Cheng, J. Wei, P. Shangguan, and W. Huang, "Emotion recognition based on multi-variant correlation of physiological signals," *IEEE Transactions on Affective Computing,* vol. 5, no. 2, pp. 126-140, 2014.

[69] A. Albraikan, D. P. Tobón, and A. El Saddik, "Toward user-independent emotion recognition using physiological signals," *IEEE Sensors Journal,* vol. 19, no. 19, pp. 8402-8412, 2018.

[70] O. Bălan, G. Moise, L. Petrescu, A. Moldoveanu, M. Leordeanu, and F. Moldoveanu, "Emotion classification based on biophysical signals and machine learning techniques," *Symmetry,* vol. 12, no. 1, p. 21, 2019.

[71] Y. Liu, T. Gedeon, S. Caldwell, S. Lin, and Z. Jin, "Emotion Recognition Through Observer's Physiological Signals," *arXiv preprint arXiv:2002.08034,* 2020.

[72] G. Oh *et al.*, "Drer: Deep learning–based driver's real emotion recognizer," *Sensors,* vol. 21, no. 6, p. 2166, 2021.

[73] J. T. Cacioppo and L. G. Tassinary, "Inferring psychological significance from physiological signals," *American psychologist,* vol. 45, no. 1, p. 16, 1990.

[74] C. M. Jones and T. Troen, "Biometric valence and arousal recognition," in *OzCHI*, 2007 2007, pp. 191-194.

[75] Z. Khalili and M. H. Moradi, "Emotion detection using brain and peripheral signals," in *CIBEC*, 2008 2008: IEEE, pp. 1-4.

[76] Y. Gu, S. L. Tan, K. J. Wong, M. H. R. Ho, and L. Qu, "Using GA-based feature selection for emotion recognition from physiological signals," in *2008 International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS2008)*, 2009 2008: IEEE, pp. 1-4.

[77] P. A. Nogueira, R. Rodrigues, E. Oliveira, and L. E. Nacke, "A regression-based method for lightweight emotional state detection in interactive environments," in *XVI Portuguese conference on artificial intelligence (EPIA)*, 2013 2013.

[78] C. A. Torres-Valencia, H. F. Garcia-Arias, M. A. A. Lopez, and A. A. Orozco-Gutiérrez, "Comparative analysis of physiological signals and electroencephalogram (EEG) for multimodal emotion recognition using generative models," in *2014 XIX Symposium on Image, Signal Processing and Artificial Vision*, 2014 2014: IEEE, pp. 1-5.

[79] M. B. H. Wiem and Z. Lachiri, "Emotion classification in arousal valence model using MAHNOB-HCI database," *International Journal of Advanced Computer Science and Applications*, vol. 8, no. 3, 2017.

[80] M. B. H. Wiem and Z. Lachiri, "Emotion sensing from physiological signals using three defined areas in arousal-valence model," in *2017 International Conference on Control, Automation and Diagnosis (ICCAD)*, 2017 2017: IEEE, pp. 219-223.

[81] P. Kawde and G. K. Verma, "Deep belief network based affect recognition from physiological signals," in *4th IEEE Uttar Pradesh Section International Conference on electrical, computer and electronics (UPCON)*, 2017 2017: IEEE, pp. 587-592.

[82] W. M. B. Henia and Z. Lachiri, "Emotion classification in arousal-valence dimension using discrete affective keywords tagging," presented at the International Conference on Engineering & MIS (ICEMIS), 2017, 2017.

[83] E. J. Choi and D. K. Kim, "Arousal and valence classification model based on long short-term memory and deap data for mental healthcare management," *Healthcare informatics research*, vol. 24, no. 4, pp. 309-316, 2018.

[84] S. Sarabadani, L. C. Schudlo, A. A. Samadani, and A. Kushski, "Physiological detection of affective states in children with autism spectrum disorder," *IEEE Transactions on Affective Computing*, vol. 11, no. 4, pp. 588-600, 2018.

[85] M. Ali, F. Al Machot, A. Haj Mosa, M. Jdeed, E. Al Machot, and K. Kyamakya, "A globally generalized emotion recognition system involving different physiological signals," *Sensors*, vol. 18, no. 6, p. 1905, 2018.

[86] D. Ayata, Y. Yaslan, and M. E. Kamasak, "Emotion based music recommendation system using wearable physiological sensors," *IEEE transactions on consumer electronics*, vol. 64, no. 2, pp. 196-203, 2018.

[87] A. Albraikan, D. P. Tobón, and A. El Saddik, "Toward user-independent emotion recognition using physiological signals," *IEEE sensors Journal*, vol. 19, no. 19, pp. 8402-8412, 2019.

[88] C. Li, Z. Bao, L. Li, and Z. Zhao, "Exploring temporal representations by leveraging attention-based bidirectional LSTM-RNNs for multi-modal emotion recognition," *Information Processing & Management*, vol. 57, no. 3, p. 102185, 2020.

[89] M. Baghizadeh, K. Maghooli, F. Farokhi, and N. J. Dabanloo, "A new emotion detection algorithm using extracted features of the different time-series generated from ST intervals Poincaré map," *Biomedical Signal Processing and Control*, vol. 59, p. 101902, 2020.

[90] "The All Music Guide." [Online]. Available: https://www.allmusic.com/

[91] B.-j. Han, S. Rho, R. B. Dannenberg, and E. Hwang, "SMERS: Music Emotion Recognition Using Support Vector Regression," in *ISMIR*, 2009: Citeseer, pp. 651-656.

[92] M. Soleymani, S. Koelstra, I. Patras, and T. Pun, "Continuous emotion detection in response to music videos," in *IEEE International Conference on Automatic Face & Gesture Recognition (FG)*, 2011: IEEE, pp. 803-808.

[93] C. A. Torres-Valencia, M. A. Álvarez, and Á. A. Orozco-Gutiérrez, "Multiple-output support vector machine regression with feature selection for arousal/valence space emotion assessment," in *36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2014 2014: IEEE, pp. 970-973.

[94] M. E. Oswald and S. Grosjean, "Confirmation bias," *Cognitive illusions: A handbook on fallacies and biases in thinking, judgement and memory*, vol. 79, p. 83, 2004.

[95] M. F. King and G. C. Bruner, "Social desirability bias: A neglected aspect of validity testing," *Psychology & Marketing*, vol. 17, no. 2, pp. 79-103, 2000.

[96] K. Sharma, C. Castellini, F. Stulp, and E. L. Van den Broek, "Continuous, real-time emotion annotation: A novel joystick-based analysis framework," *IEEE Transactions on Affective Computing*, vol. 11, no. 1, pp. 78-84, 2017.

[97] P. Van Gent, H. Farah, N. Van Nes, and B. Van Arem, "HeartPy: A novel heart rate algorithm for the analysis of noisy signals," *Transportation research part F: traffic psychology and behaviour*, vol. 66, pp. 368-378, 2019.

[98] R. Satti *et al.*, "The application of the extended Poincaré plot in the analysis of physiological variabilities," *Frontiers in physiology*, vol. 10, p. 116, 2019.

[99] W. Karlen, S. Raman, J. M. Ansermino, and G. A. Dumont, "Multiparameter respiratory rate estimation from the photoplethysmogram," *IEEE Transactions on Biomedical Engineering*, vol. 60, no. 7, pp. 1946-1953, 2013.

[100] M. Waskom. "seaborn 0.11.2." https://seaborn.pydata.org/ (accessed.

[101] M. Weber. "statannot 0.2.3." https://github.com/webermarcolivier/statannot (accessed.

[102] H. Han, W.-Y. Wang, and B.-H. Mao, "Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning," 2005: Springer, pp. 878-887.

[103] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *the Journal of machine Learning research*, vol. 12, pp. 2825-2830, 2011.

[104] x. developers. "xgboost." https://xgboost.readthedocs.io/en/stable/python/index.html (accessed.

[105] F. a. o. Chollet. "Keras." https://keras.io (accessed.

[106] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, 2012.

[107] N. Smith, "Galvanic Skin Sensor technology on wearables is just getting started." [Online]. Available: https://www.myhealthyapple.com/galvanic-skin-sensor-technology-on-wearables-is-just-getting-started/

[108] D. Watson and K. Stanton, "Emotion blends and mixed emotions in the hierarchical structure of affect," *Emotion Review*, vol. 9, no. 2, pp. 99-104, 2017.

APPENDICES

APPENDIX A: FULL CLASSIFICATION RESULTS

TABLE 26. 5-FOLD CROSS VALIDATION ACCURACY FOR EACH MODEL AND FEATURE SET

	Neural Net		Random Forest		SVM		1D CNN		LSTM	
Feature Set	Accuracy	Accuracy SD	Accuracy	Accuracy SD	Accuracy	Accuracy SD	Accuracy	Accuracy SD	Accuracy	Accuracy SD
No A/V (Only Physiological)	0.4844	0.0508	0.6534	0.0240	0.6087	0.0346	0.3692	0.0502	0.3652	0.0684
True A/V (CASE Dataset) + Physiological	0.9178	0.0028	0.9475	0.0038	0.8766	0.0031	0.9296	0.1088	0.7321	0.0493
Linear Regression A/V + Physiological	***0.9833***	0.0089	***0.9388***	0.0217	***0.9879***	***0.0029***	***0.9207***	0.0510	0.5980	0.0533
Random Forest Regression A/V + Physiological	0.6633	0.0298	0.7802	0.0087	0.6557	0.0125	0.5251	0.0430	0.5062	0.0658
SVM Regression A/V + Physiological	0.7549	0.0207	0.8224	0.0145	0.8217	0.0221	0.6505	0.0499	0.5192	0.1023
AdaBoost Regression A/V + Physiological	***0.9249***	0.0175	***0.9667***	0.0147	0.8574	0.0181	0.7879	0.0517	0.6328	0.0414
XGBoost Regression A/V + Physiological	0.6398	0.0353	0.7549	0.0137	0.6471	0.0149	0.5221	0.0562	0.5044	0.0644

TABLE 27. 5-FOLD CROSS VALIDATION AUC FOR EACH MODEL AND FEATURE SET

	Neural Net		Random Forest		SVM		1D CNN		LSTM	
Feature Set	AUC	AUC SD	AUC	AUC SD	AUC	AUC SD	AUC	AUC SD	AUC	AUC SD
No A/V (Only Physiological)	0.7542	0.0391	0.8856	0.0165	0.7554	0.0216	0.6559	0.0515	0.6792	0.0685
True A/V (CASE Dataset) + Physiological	0.9910	0.0010	0.9962	0.0007	0.9229	0.0019	0.9783	0.0301	0.9303	0.0296
Linear Regression A/V + Physiological	***0.9990***	0.0011	***0.9963***	0.0015	***0.9994***	***0.0002***	***0.9888***	0.0116	0.8795	0.0189
Random Forest Regression A/V + Physiological	0.8912	0.0174	0.9593	0.0061	0.7848	0.0078	0.7960	0.0353	0.7798	0.0589
SVM Regression A/V + Physiological	0.9381	0.0088	0.9700	0.0044	0.8886	0.0138	0.8724	0.0388	0.7913	0.0859
AdaBoost Regression A/V + Physiological	***0.9880***	0.0060	***0.9992***	0.0005	0.9109	0.0113	0.9286	0.0308	0.8826	0.0411
XGBoost Regression A/V + Physiological	0.8794	0.0184	0.9501	0.0079	0.7794	0.0093	0.7788	0.0496	0.7936	0.0578

TABLE 28. 5-FOLD CROSS VALIDATION F1 SCORE FOR EACH MODEL AND FEATURE SET

	Neural Net		Random Forest		SVM		1D CNN		LSTM	
Feature Set	F1 Score	F1 Score SD	F1 Score	F1 Score SD	F1 Score	F1 Score SD	F1 Score	F1 Score SD	F1 Score	F1 Score SD
No A/V (Only Physiological)	0.4551	0.2066	0.6111	0.2241	0.5772	0.2461	0.3341	0.2186	0.3339	0.2251
True A/V (CASE Dataset) + Physiological	0.9180	0.0539	0.9474	0.0372	0.8760	0.1169	0.9244	0.1388	0.7113	0.2244
Linear Regression A/V + Physiological	***0.9833***	0.0152	***0.9370***	0.0614	***0.9879***	***0.0098***	***0.9150***	0.1302	0.5595	0.2864
Random Forest Regression A/V + Physiological	0.6492	0.1583	0.7643	0.1446	0.6386	0.1656	0.5024	0.2209	0.4835	0.2215
SVM Regression A/V + Physiological	0.7502	0.1396	0.8125	0.1279	0.8169	0.1040	0.6309	0.2291	0.4969	0.2392
AdaBoost Regression A/V + Physiological	***0.9245***	0.0566	***0.9662***	0.0365	0.8561	0.0925	0.7748	0.2146	0.6110	0.2516
XGBoost Regression A/V + Physiological	0.6269	0.1467	0.7373	0.1493	0.6297	0.1615	0.4977	0.2145	0.4802	0.2073

TABLE 29. 5-FOLD CROSS VALIDATION PRECISION FOR EACH MODEL AND FEATURE SET

	Neural Net		Random Forest		SVM		1D CNN		LSTM	
Feature Set	Precision	Precision SD	Precision	Precision SD	Precision	Precision SD	Precision	Precision SD	Precision	Precision SD
No A/V (Only Physiological)	0.4470	0.1537	0.6283	0.1050	0.5757	0.1797	0.3512	0.1902	0.3592	0.1943
True A/V (CASE Dataset) + Physiological	0.9223	0.0704	0.9483	0.0395	0.8765	0.1134	0.9496	0.1088	0.7281	0.1893
Linear Regression A/V + Physiological	***0.9833***	0.0152	***0.9437***	0.0603	***0.9882***	***0.0160***	***0.9358***	0.0980	0.6538	0.2688
Random Forest Regression A/V + Physiological	0.6734	0.1322	0.7940	0.0969	0.6564	0.1284	0.5376	0.2106	0.5196	0.2311
SVM Regression A/V + Physiological	0.7753	0.1438	0.8340	0.0965	0.8376	0.1050	0.6510	0.2096	0.5344	0.2617
AdaBoost Regression A/V + Physiological	***0.9309***	0.0714	***0.9690***	0.0408	0.8663	0.1017	0.8055	0.1770	0.6432	0.2255
XGBoost Regression A/V + Physiological	0.6440	0.1179	0.7662	0.0919	0.6476	0.1207	0.5292	0.2044	0.5258	0.2064

TABLE 30. 5-FOLD CROSS VALIDATION RECALL FOR EACH MODEL AND FEATURE SET

	Neural Net		Random Forest		SVM		1D CNN		LSTM	
Feature Set	**Recall**	**Recall SD**	**Recall**	**Recall SD**	**Recall**	**Recall SD**	**Recall**	**Recall SD**	**Recall**	**Recall SD**
No A/V (Only Physiological)	0.4844	0.2671	0.6534	0.2980	0.6087	0.2829	0.3652	0.2854	0.3610	0.2927
True A/V (CASE Dataset) + Physiological	0.9178	0.0656	0.9476	0.0458	0.8766	0.1226	0.9307	0.1737	0.7311	0.2787
Linear Regression A/V + Physiological	***0.9833***	0.0220	***0.9388***	0.0972	***0.9879***	***0.0098***	***0.9213***	0.1649	0.5974	0.3352
Random Forest Regression A/V + Physiological	0.6633	0.2182	0.7802	0.2159	0.6557	0.2279	0.5223	0.2808	0.5031	0.2811
SVM Regression A/V + Physiological	0.7549	0.1803	0.8224	0.1881	0.8217	0.1559	0.6489	0.2753	0.5162	0.2773
AdaBoost Regression A/V + Physiological	***0.9249***	0.0795	***0.9667***	0.0609	0.8574	0.1208	0.7876	0.2485	0.6314	0.2973
XGBoost Regression A/V + Physiological	0.6398	0.2080	0.7549	0.2213	0.6471	0.2275	0.5194	0.2810	0.5014	0.2703

VITA

Matthew Nathanael Gray

ODU Department of Electrical and Computer Engineering
231 Kaufman Hall, Norfolk, VA 23529
Email: mgray564@gmail.com
LinkedIn: https://www.linkedin.com/in/matthewngray/

Education

MS in Electrical and Computer Engineering – Old Dominion University
Norfolk, VA — 2016-2022

BS in Biomedical Engineering – University of Alabama at Birmingham
Birmingham, AL — 2012-2016

Work Experience

Electrical and Computer Engineer – NASA Langley Research Center
Hampton, VA — June 2020 - Present

Graduate Research Assistant – ODU Virginia Modeling, Analysis, and Simulation Center (VMASC)
Suffolk, VA — Sept 2019 – May 2020

Pathways Co-op – NASA Armstrong Flight Research Center
Edwards, CA — May 2019 – Aug 2019

Pathways Co-op – NASA Armstrong Flight Research Center
Edwards, CA — January 2018 - August 2018

Intern – NASA Langley Research Center
Hampton, VA — June 2017- August 2017

Intern – NASA Langley Research Center
Hampton, VA — January 2017-May 2017

Graduate Research Assistant – Advanced Signal Processing in Engineering and Neuroscience (ASPEN) Lab
Old Dominion University – Norfolk, VA – August 2016-May 2018

Intern – NASA Langley Research Center
Hampton, VA — June 2016-August 2016

Undergraduate Research Assistant – Dr. Amthor's Lab
University of Alabama at Birmingham – Birmingham, AL — October 2015-May 2016

Intern – NASA Langley Research Center
Hampton, VA — May 2015-August 2015

Honors/Awards

Steven B. Davis Co-op/Student Award - NASA Armstrong Center-wide Peer Award
NASA Armstrong Flight Research Center – Edwards, CA – November 2019

Eagle Scout
Troop 71, Birmingham, AL – May 2012

Golden Excellence Scholarship
University of Alabama at Birmingham — 2012-2016

Presentations/Publications

Annual BMES Conference Platform Presentation
Minneapolis, Minnesota — Oct. 2016