Multi-modal Emotional Recognition using ECG and GSR

Ryan Smyth

rsmyth1@gmail.com

1. INTRODUCTION

Emotions are a constant across the human experience. While emotions can be shaped and molded by each person's cultural context, there are universal, so-called basic, emotions recognized as far back as French philosopher and scientist Descartes (Descartes 1649). Darwin recognized the importance of these basic emotions for survival and highlighted the importance of accurate emotional recognition (Darwin 1872). In today's modern world, emotional recognition is not typically a matter of survival, but it does represent an important field as we try to expand this natural human ability to the technology we wield every day.

Emotional recognition can be beneficial in many fields, and the complicated interplay of several fields such as psychology, neuroscience, and computer science has given rise to the field of affective computing. More directly, emotional recognition has helped facilitate healthcare outcomes (Guo et al. 2024) and seems to aid academic achievement in children (Bulut 2018). The potential benefits have led to a growing number of studies examining not only the benefits of emotional recognition, but how to facilitate computer recognition of human emotions in the first place.

One chief challenge to emotional recognition is the quantification of emotions. To that end, the Circumplex model, containing two axes that work together to characterize emotions, was developed by Russell (1980). The first axis is emotional valence, or whether the emotion is positive or negative. The other axis is emotional arousal, which records the magnitude of emotional response. For example, joy is high on both valence and arousal, content is high on valence, low on arousal, rage is low on valence and high on arousal, and sadness is low on both valence and arousal.

1.1 Related work

Many studies have been published regarding systems and programs to detect and classify human emotions. Initially information from observation, e.g. photographs, was used,

but as the field grew, the rise of using physiological data collected by instruments has become more popular given its perceived reliance and objectivity (Pan, Hirota, Jia & Dai 2023). Researchers have more recently considered how humans recognize emotions and the multiple signals we receive and process. As a result they have worked on combining different signals to determine if the fusion of signals can improve accuracy and lead to better predictions (Ezzameli and Mahersia 2023). This drive for better predictions and fusion of signals has led to the rapid growth of multi-modal emotion recognition (MER), however a key limitation appears to persist.

Many studies rely on a combination of technologies that are impractical for *in vivo* use. For instance many rely on data gathered with multi-lead electroencephalogram headsets as seen in *Figure 1*, or continue to include observational data such as video and audio recordings which would introduce significant amounts of noise in non-controlled settings.

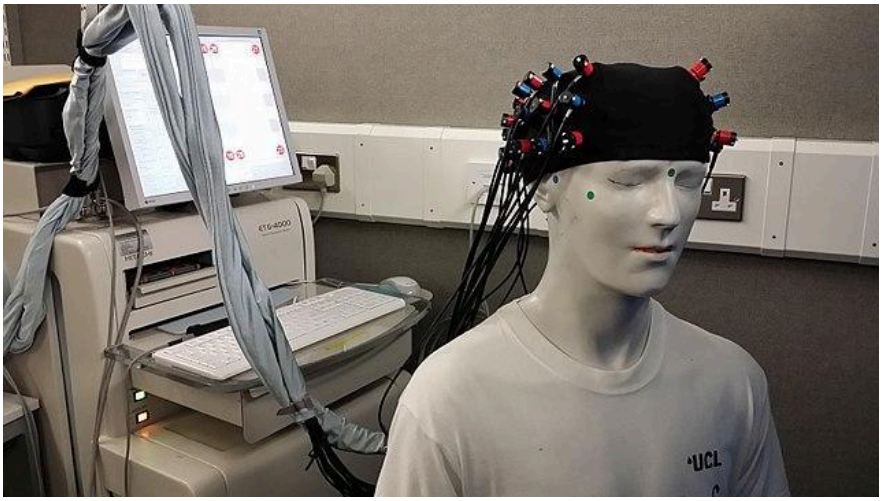

Figure 1 - An example electroencephalogram headset

1.2 This work

The aim of this project is to address the limitations found in existing studies by taking advantage of the development of wearable health monitoring technology. By using signals easily gathered using common tools such as a smart watch, smart ring, or other common wearable devices, MER can move closer to practical use in everyday situations. To that end this project evaluated 2 physiological signals often associated with changes in stress levels and the emotional state: electrocardiogram (ECG) and galvanic skin response (GSR), also known as skin conductivity. These signals are able to be captured by modern smart watches and will allow the

exploration of MER using a more practical data source to classify emotions using several machine learning techniques including multilayer perceptrons, random forests, logistic regression classifier, ridge regression classifier, support vector machines, and a classifier using random forests combined with the adaboost algorithm. The project evaluates children, but could serve as proof of concept for other populations as well.

2. METHODOLOGY

2.1 The data set

This project utilizes the Young Adult Affective Dataset (YAAD). Not every participant in the full dataset underwent both ECG and GSR recording. When considering only participants who recorded both ECG and GSR, the reduced dataset consists of 12 participants in the age range of 8 - 14. These participants watched emotionally charged video clips and filled out self response forms regarding their emotional state following each video (Dar et al. 2022). *Figure 2* contains images about sex distribution of the participants, while *Figure 3* contains information about the age breakdown of participants. Participants rated their level of felt emotion in the basic emotions of happy, sad, fear, surprise, anger, disgust, and additionally a neutral category. Participants also rated their emotional valence, whether the emotion was positive or negative, and emotional arousal, severity/magnitude of emotional reaction.

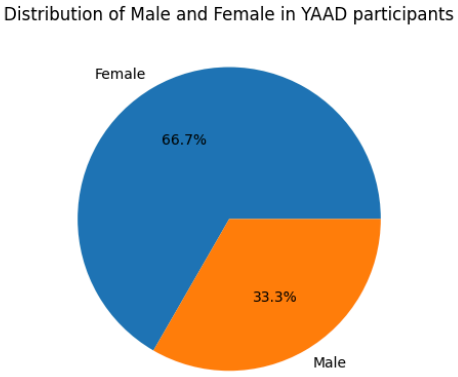

Figure 2 - Sex distribution of participants

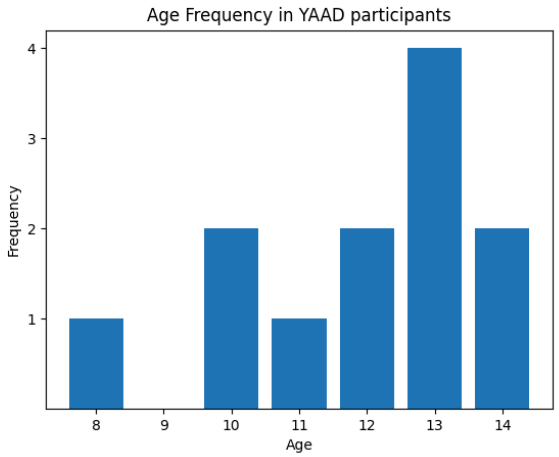

Figure 3 - Age distribution of participants

Each participant had 3 sessions and in each session watched 7 videos, one for each of the basic emotions. Each video consists of two components, a 5 second lead in period designed to elicit a neutral response and a 34 second long period where emotionally charged material is presented. When considering the data, this neutral lead in period was removed leaving only physiological data associated with the emotional portion of the video. A gap was given to participants between videos to allow a return to a neutral emotional state. Considering each video viewing as a trial led to a total of 252 individual trials as the data used in this project.

Sampling was taken using Shimmer3 ECG and GSR units worn on the chest and palm respectively. The sampling rate was 256 Hz which resulted in 5000 data points. Steps were taken to minimize noise by removing subject movement and removing unnecessary electronic devices from the room.

The record of self-reporting responses has 17 columns; 3 columns are ID values designed to identify the participant, session, and video being watched. 3 columns are demographic information of age, sex, and participant initials. 3 columns then report the valence level, arousal level, and dominance level on a scale ranging from 0 to 10. 7 columns, one for each emotion, follow with responses ranging from “very low” to “very high”. Finally, a familiarity score was taken to assess how familiar an individual is with the video being presented.

2.2 Response variables

This project ultimately attempted to explore predictions of 4 response variables. The first and second predictions were for valence and arousal levels independently. For this, each score from 0 to 10 was categorized as "low" if falling in 0-4 range, or "high" if falling in 5-10 range. This categorization of the valence and arousal scores were represented as "LV" and "HV" for low and high valence respectively. Similarly, arousal was represented by "LA" and "HA" for low and high arousal respectively.

The third predicted value is a combination of valence and arousal to get the emotion's location on the full valence-arousal axis. This response is generated simply by combining the possible categories of the valence and arousal responses to create 4 categories "HVHA" for high in both domains, "HVLA" for high in valence and low in arousal, "LVHA" for low in valence and high in arousal, and "LVLA"for low scores in both domains.

The final was a prediction of the strongest emotion reported by each participant. To do this, each emotion's rating of "very low", "low", "moderate", "high", and "very high" were changed to numeric values 1, 2, 3, 4, and 5 respectively. The column with the highest value was taken to be the dominant emotion and that column's label was made the target variable giving happy, sad, fear, surprise, anger, disgust, and neutral as possible responses.

2.3 Signal filtering

2.3.1 *ECG signal filtering*

ECG signals can be prone to noise caused by minor movements, electrical devices in the room, and noise signals generated from the powerline used by the device. As a result a butterworth passband filter was used to remove signals below 0.03 Hz and above 60 Hz. A notch filter was used to remove the 50 Hz power line noise signal.

2.3.2 *GSR signal smoothing*

GSR does not have external sources of noise, however it is susceptible to fluctuations caused by the body's electrical activity, such as minor or involuntary muscle movements. As a result the filtering required is a smoothing process designed to eliminate the rapid fluctuation of background bodily activity. A 3rd order Savgol filter with a window length of 15 was used to smooth the signal as seen in *Figure 4.*

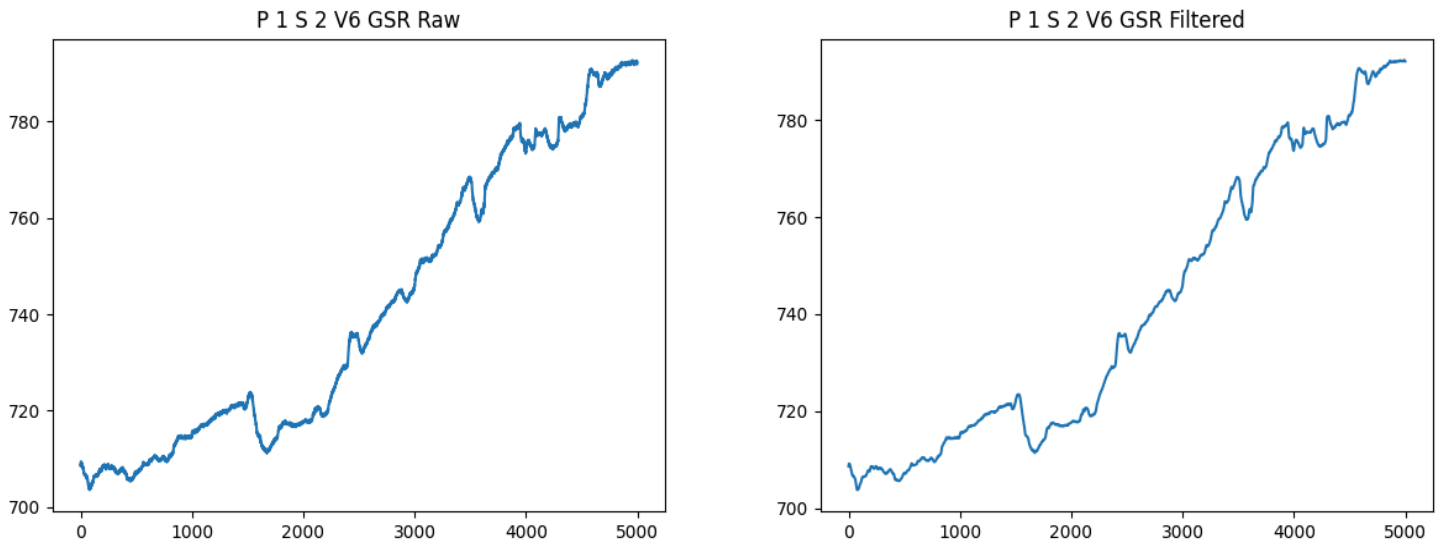

Figure 4 - GSR Smoothing Results

2.4 Feature Extraction

2.4.1 *ECG time domain feature extraction*

This project utilized two main time domain approaches to extract features from the ECG signal. The first is gathering information on metrics associated with heart rate variability (HRV) as discussed in Rainville, Bechara, Naqvi & Damasio (2006). In order to carry out analysis on the HRV of each trial, the component of the QRS complex found in ECG signals dubbed the R peak needed to be identified. These R peaks were identified as seen in *Figure 5* using the *neurokit2* python library and the Emrich 2023 detection algorithm (Makowski et al. 2021).

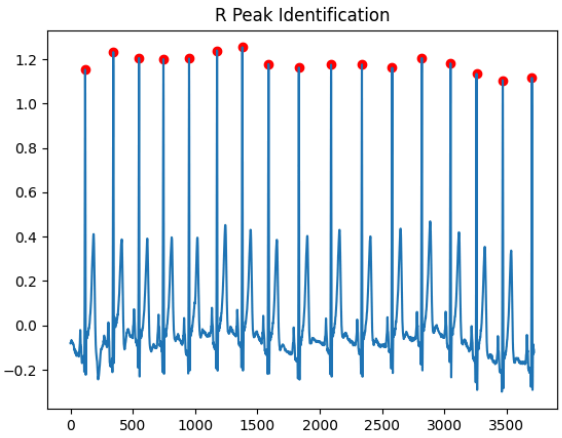

Figure 5 - R peak identification using Neurokit2 library

Once the R peaks were identified 4 features were extracted from the dataset. The interval between R peaks was used to generate 4 features that provide information about HRV. *Table 1* contains a brief description of the features associated with using the R-R interval to look at HRV.

Table 1—HRV associated features

Name	Description
R-R Interval Mean	The average time between R-R peaks
R-R Interval Standard Deviation	The standard deviation of the R-R intervals
Root Mean Squared Differences	Measures variability in heart rate
pNN50	The percentage of R-R intervals that are greater than 50ms

In addition to HRV, other information about the ECG signal was gathered. This includes general statistics about the ECG signal such as the mean and standard deviation. Kurtosis, skew, and variance of the signal was also calculated and used as features.

2.4.2 *ECG Frequency domain feature extraction*

In addition to the time domain analysis, steps were taken to look at the frequency domain of the signal. The mean, standard deviation, kurtosis, skewness, and variance were calculated for the frequency. From here, two main approaches were tested. First, a fast Fourier transformation (FFT) which helped to inform the frequency distribution of the signal as seen in *Figure 6.*

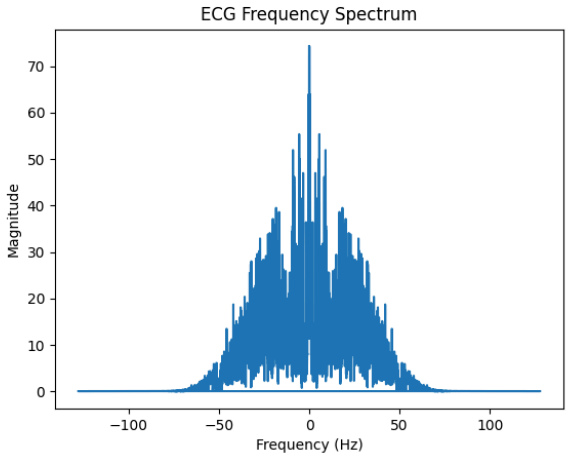

Figure 6 - Frequency spectrum of one participant's ECG

Once the frequency distribution was determined via FFT, several features were extracted from this information. For example, bands of high and low frequencies were identified; the bands with low frequencies are associated with increased stress/negative emotions and high frequencies are associated with a more relaxed/positive state (Malliani 1993). This identification of high vs low frequencies can be seen in *Figure 7.*

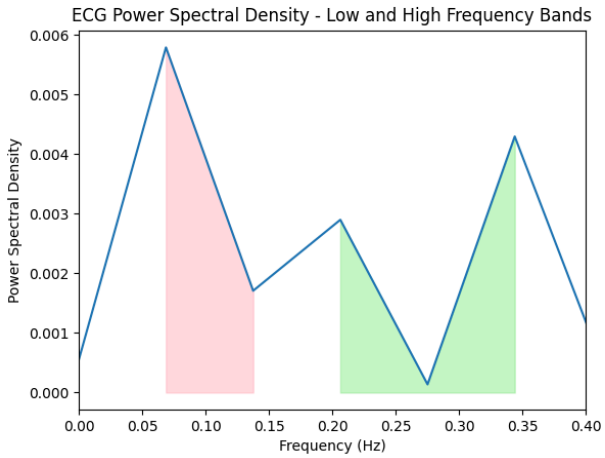

Figure 7 - The identification of high frequency (green) vs. low frequency (red)

However, once DWT was included for GSR, as discussed in the next section, DWT was also explored with ECG to evaluate the frequency changes over time. The features extracted using these transforms can be seen in *Table 2* for the FFT and *Table 3* for the DWT.

Table 2—Fast fourier transform associated features

Name	Description
Spectral Centroid	The average frequency power
High-Low Power Ratio	The ratio of power in high frequency bands compared to power in low frequency bands.
Low Frequency Skew	Measures variability in heart rate
HIgh Frequency Skew	The percentage of R-R peaks that are greater than 50ms
Low Frequency Range	The range of power found in low power frequencies.
High Frequency Range	The range of power found in high power frequencies.

Table 3—Daubechies wavelet associated features

Name	Description
Energy	The magnitude of power in the coefficients.
Mean	The average of the coefficients
Standard Deviation	The standard deviation of the coefficients
Maximum Coefficient	The largest value found in this coefficient value
Minimum Coefficient	The lowest coefficient value
Entropy	Entropy of the Signal

2.4.3 *GSR feature extraction*

Initially, features extracted from GSR were limited to the general descriptive statistics, however, poor accuracy and limited information gained this way prompted the generation of features from the frequency domain as well. Ultimately, features extracted from the GSR include information similar to those extracted from the ECG. In the time domain, mean, standard deviation, kurtosis, skew, and variance were extracted. Initially the range of the GSR was also included, however this caused a decrease in accuracy. In an attempt to get a more complete picture of the GSR signal, more information from the frequency domain was included. In order to get this information from the frequency domain, DWT was used to extract the features seen in *Table 3.*

2.5 Data preparation and fusion

This project aimed to explore the differences in efficacy of both early and late fusion. To achieve this, some data preparation steps were taken first. First, the data of the two signals was scaled using sk-learn's minmax scaler, and then the data was divided into 70% training data and 30% testing data. Once the data was split, sk-learn's SelectKBest function was used to select the 10 features most associated with the prediction tag. From here the data was handled in two different ways. First was early fusion, where ECG and GSR data was joined and the combined feature array was input directly to the models. The second was that the ECG and GSR data was input separately to models before being used in the late fusion process described below.

To explore late fusion, this project utilized a "majority vote" system. 5 models were trained on the ECG data, and another 5 were trained on the GSR data. Each model made a prediction for the testing data and the response most common among the 10 models was taken as the predicted response. This was done for all 4 prediction categories. Given the data fed to each of the models would be the same, deterministic models, such as logistic regression, ridge classifier, and support vector machines, were excluded from this step as all the models would return the same values.

2.6 Identifying the most successful transforms

Several parameters needed to be explored to determine the best accuracy possible. To achieve this, the numpy library's GridSearchCV function was used to explore hyperparameters and different transforms and wavelets for ECG and GSR. This led to a 3 step process. The first step was exploring the transforms for the GSR data, the second step was exploring transforms for ECG, and the final step was utilizing the best transforms with hyperparameter tuning to find the best performing models.

The first exploration was considering what wavelets to utilize with GSR and ECG. To do so, ECG was held constant by using FFT, while GSR was analyzed while fused to ECG data (representing early fusion), and by itself (representing the prediction it would undergo during late fusion). Using the Pywavelets library, wavelets db2 and db5 were tested at levels 1 through 5 for multilayer perceptron, random forest, logistic regression, ridge classifier, support vector machine, and adaboost algorithm using random forest classifiers. The average accuracy of the models was then taken for each of the four target tags: valence, arousal, the full valence-arousal axis, and emotional tags. This average accuracy was used to find the strongest consistent performance.

Once the highest performing transforms were found for the GSR signal, a similar search was carried out to analyze the performance of the Daubechies transforms and fast Fourier transform for the ECG signal. Similar to the analysis of transforms for GSR, each of the six models was run with the data and the mean performance was used to find the consistently strongest performance among the transforms for the four prediction targets. This was done in the case of early fusion, where ECG data was fused to the GSR data generated using transforms found in the previous step, and looking at ECG data alone which would represent its inclusion in late fusion.

2.7 Final Analysis

Finally, with the best transforms for GSR and ECG found, a grid search was carried out using features extracted with these wavelets. The parameters tested for each model can be found in *Appendix 1.* First the models were trained using data generated according to the best transforms. Then the models were fit to the data and used to predict the 4 prediction categories. The models were first fed the fused ECG and GSR data to get results for the early fusion

condition. Then the majority vote system began with copies of the models receiving either ECG only or GSR only and making predictions for the four categories. As a baseline, these predictions were kept to compare multimodal methods to unimodal methods. Once these predictions were acquired, they were then fed into a voting system as previously discussed.

Finally, a statistical analysis was carried out to determine if the usage of fused data was superior to unimodal data. A two proportion Z-test was conducted comparing the accuracy of the early fused data to the accuracy of models using a single physiological signal. The equation for a two proportion Z-test can be found in *Figure 8.* A p-value of 0.05 was used for a significance threshold.

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\bar{p} \times \bar{q}}{n_1} + \frac{\bar{p} \times \bar{q}}{n_2}}}$$

Figure 8 - Two proportion Z test

3. RESULTS

3.1 Analysis of DWT for GSR

The first exploration of wavelets to use with GSR show the strongest performance with early fusion data was found using the following: db2 wavelet at level 3 for predicting valence, db2 wavelet at level 2 for predicting arousal, db2 wavelet at level 2 for predicting the full valence-arousal axis, and the db2 wavelet at level 1 for predicting the emotional tags. The average results for each transform can be seen in *Table 4.*

Table 4—Average percent accuracy for GSR DWT wavelets when models trained with ECG and GSR data fused

Wavelet	Level	Valence Accuracy	Arousal Accuracy	VA Axis Accuracy	Emotion Accuracy
DB2	1	70.68	71.24	55.26	55.26
	2	70.49	76.50	56.20	50.56

	3	76.50	68.23	51.50	51.88
	4	72.56	61.84	45.11	51.32
	5	72.74	69.74	53.20	51.50
DB5	1	65.23	66.92	48.50	42.48
	2	72.37	62.03	42.67	41.35
	3	65.98	62.78	43.61	46.80
	4	67.67	64.29	41.17	43.05
	5	60.34	58.65	39.29	38.72

When considering GSR alone, e.g. in the case of late fusion, the strongest performance was the following: db2 at level 2 for valence, db2 at level 2 for arousal, db2 at level 4 for the valence-arousal axis, and db5 at level 2 for the emotional tags. The average results for each transform can be seen in *Table 5*.

Table 5—Average percent accuracy for GSR DWT wavelets for late fusion including GSR only

Transform	**Level**	**Valence Accuracy**	**Arousal Accuracy**	**VA Axis Accuracy**	**Emotion Accuracy**
DB2	1	69.74	68.80	52.44	50.19
	2	78.57	69.36	53.20	52.82
	3	72.18	69.17	47.18	51.69
	4	76.32	66.17	53.38	46.62
	5	72.93	66.73	51.32	51.50
DB5	1	75.94	62.78	51.69	48.12
	2	77.07	68.98	51.13	53.95
	3	77.07	63.53	46.99	49.44
	4	75.56	62.41	52.82	45.30

	5	73.12	64.66	41.17	49.62

3.2 Analysis of DWT and FFT for ECG

The results of the second step of the analysis, analyzing the wavelets for ECG, show the strongest performance was as follows: db2 wavelet at level 1 for valence, db2 at level 4 for arousal, the db2 wavelet at level 1 for the valence-arousal axis, and db5 at level 3 for the emotional tags. The averages for each transform can be seen in *Table 6.*

Table 6—Average percent accuracy for ECG DWT wavelets and FFT for early fusion including both ECG and GSR.

Transform	Level	Valence Accuracy	Arousal Accuracy	VA Axis Accuracy	Emotion Accuracy
DB2	1	79.14	65.23	52.63	48.68
	2	75.00	64.47	48.87	47.37
	3	74.62	67.67	50.19	46.99
	4	76.32	70.86	52.44	49.06
	5	64.10	69.92	43.23	45.11
DB5	1	64.66	64.47	43.80	50.94
	2	74.81	64.10	45.86	44.92
	3	71.80	66.92	50.94	53.01
	4	73.12	61.47	49.44	50.56
	5	64.47	65.04	45.30	46.43
FFT	-	74.62	65.23	51.13	51.69

The full results for the analysis of ECG alone, representing a step of the late fusion approach, indicate the strongest performance was seen by db5 wavelet at level 2 for valence, db2 at level 5 for arousal, db2 at level 4 for the valence-arousal axis, and the FFT for the emotional tags. The averages for each transform can be seen in *Table 7.*

Table 7—Average percent accuracy for ECG DWT wavelets for late fusion including ECG only

Wavelet	Level	Valence Accuracy	Arousal Accuracy	VA Axis Accuracy	Emotion Accuracy
DB2	1	68.23	66.17	44.17	46.05
	2	66.92	66.92	47.18	44.74
	3	65.41	63.91	39.47	42.11
	4	68.42	67.86	48.87	49.81
	5	57.33	72.56	41.92	45.49
DB5	1	57.14	64.85	34.02	47.56
	2	70.68	62.22	44.92	39.47
	3	68.80	66.73	44.17	49.25
	4	69.55	61.47	40.04	48.12
	5	57.33	66.73	37.59	43.42
FFT	-	65.23	65.79	47.56	51.50

3.3 Analysis of prediction of single physiological signal

Considering this project aims to explore the viability of MER it can be beneficial to have a comparison to unimodal results. Since the late fusion approach of this project involves using the two signals independently, results for predicting the tags with only ECG and GSR can be found below in *Table 8* and *Table 9* respectively.

Table 8—Percent accuracy when predicting from only ECG

Model	Valence Accuracy	Arousal Accuracy	VA Axis Accuracy	Emotion Accuracy
Multilayer Perceptron	68.42	69.74	47.37	47.37

Random Forest	76.32	68.42	50.00	52.63
Logistic Regression	68.42	64.47	38.16	44.74
Ridge Classifier	68.42	67.11	34.21	44.74
K Nearest Neighbors	63.16	71.05	46.05	43.42
Support Vector Classifier	69.74	60.53	38.16	51.32
Adaboost Classifier	71.05	67.11	43.42	51.32

Table 9—Percent accuracy when predicting from only GSR

Model	**Valence Accuracy**	**Arousal Accuracy**	**VA Axis Accuracy**	**Emotion Accuracy**
Multilayer Perceptron	77.63	65.79	55.26	47.37
Random Forest	69.74	73.68	60.53	50.00
Logistic Regression	76.32	65.79	50.00	43.42
Ridge Classifier	78.95	63.16	53.95	47.37
K Nearest Neighbors	60.53	73.68	53.95	47.37
Support Vector Classifier	78.95	68.42	53.95	51.32
Adaboost Classifier	65.79	68.42	52.63	36.84

3.4 Final results

Once the best transforms were found for ECG and GSR the final analysis was carried out. Each model made a prediction for each of the 4 target categories with ECG and GSR data already fused (early fusion). These results can be found in *Table 10.*

Table 10—Percent accuracy with early fused data

Model	Valence Accuracy	Arousal Accuracy	VA Axis Accuracy	Emotion Accuracy
Multilayer Perceptron	77.63	72.37	53.95	50.00
Random Forest	84.21	61.84	65.79	51.32
Logistic Regression	65.79	61.84	48.68	51.32
Ridge Classifier	68.42	60.53	40.79	50.00
K Nearest Neighbors	71.05	63.16	42.11	53.95
Support Vector Classifier	69.74	53.95	52.63	51.32
Adaboost Classifier	82.89	63.16	51.32	56.58

When evaluating the efficacy of a late fusion majority voting system, 10 copies of each model made predictions for each of the 4 target categories using ECG and GSR data separately, 5 using ECG and 5 using GSR. The majority vote was then used to predict each of the 4 target categories. These results can be found in *Table 11.*

Table 11—Percent accuracy when models are trained on ECG and GSR separately and then vote on final prediction

Model	Valence Accuracy	Arousal Accuracy	VA Axis Accuracy	Emotion Accuracy
Multilayer Perceptron	3.95	5.26	3.95	1.32
Random Forest	5.26	5.26	3.95	1.32
Adaboost Classifier	5.26	5.26	3.95	1.32

3.5 Statistical analysis.

A statistical analysis was carried out to compare the accuracy of models trained on fused data to models trained on only ECG or GSR. When comparing the results of predictions made from fused data and predictions made from ECG, no statistical significance was found. Similarly, comparing predictions from fused data and GSR also found no statistical significance.

4. DISCUSSION

4.1 Evaluation of early fusion and unimodal approach

Accuracy scores for all models for early fusion and predicting using ECG and GSR alone contained a pattern. All models had a higher score for predicting valence and arousal then they did for predicting the full valence-arousal axis and emotional tags. Both valence and arousal had similar accuracy to each other, while the valence-arousal axis and emotional tag predictions were of similar accuracy. This pattern suggests identifying fully realized emotions is more difficult than representing emotional building blocks, such as emotional valence and arousal. Considering the higher accuracy of recognizing emotional subcomponents, any practical attempt to implement emotional recognition technology may wish to consider if predicting fully developed emotions is critical or if predicting just valence or arousal might be sufficient.

When considering the model accuracy using early fused data, no model is superior to the others. When averaging model results across the single signal results and the early fusion results, as seen in *Appendix 2*, the random forest classifier is the most accurate, or tied for the most accurate in predicting every category except arousal. It may not be significantly more accurate, but it does appear to be reliable across trials with strong performance. It was also the only model to surpass 60% accuracy when predicting the full VA axis.

Finally it is worth noting that fusing the ECG and GSR data did not produce a significant increase in accuracy or performance compared to evaluating ECG and GSR data alone. Consequently using a multimodal methodology does not guarantee increased performance compared to unimodal methodology and the benefits of each approach should be studied further.

4.2 Comparison of early and late fusion approaches

Evaluation of early and late fusion approaches suggests that early fusion is vastly superior to the late fusion methodology attempted in this project. Late fusion consistently returned extremely low accuracy, even going as far as 1.32% accuracy in some prediction categories. In contrast early fusion consistently returned accuracy over 45% or over 60% depending on the prediction target and model. This suggests early fusion of the data is superior to the majority vote system implemented in this project for late fusion. However, it is possible that an alternative system of late fusion with different or additional signals may produce better results.

4.3 Limitations and directions for further study

A key limitation of this study stems from the variability and limited sample size of the data set. The data set contains a limited age range, from 8 to 14, but it is an age range that might still be impacted by puberty onset. Having participants within the range of puberty onset raises questions about the physiological responses. It is possible that participants at different stages of puberty may have more or less sensitive responses. It also raises questions about response to stimuli and emotional expression at different ages. Would an 8 year old respond to all the shown videos with the same emotions as a 14 year old? This difference in emotional reaction and reactivity could lead to differences in physiological signals that create noise when trying to predict emotions. With few participants at each age point, this noise can become even more pronounced. The few participants makes outliers a greater concern as well. If one individual responded to a video with an emotion unintended by the original researchers, it could skew the data and make accurate prediction harder. Future study would benefit from a larger, more robust dataset.

Another challenge to overcome is differentiating emotions that prompt similar physiological states. For instance, heart rate may increase with a variety of emotions, such as rage and joyous surprise. This emotional differentiation challenge is a likely cause for the reduced accuracy of predicting the full VA axis and the emotional tags. Given that the accuracy of these two predictions was still better than the chance results of 25% and 14% for the VA axis and emotional tags respectively, differentiation is possible. It is also possible adding additional

data may provide more information that models can use to make this differentiation easier and lead to increased accuracy.

One area of future study is to further compare the differences between unimodal emotional recognition and MER. This project suggests they are comparable, and so more research should be done to determine key differences between these methodologies. Unimodal methodologies are naturally easier and more cost efficient, so if performance can be comparable they should be explored. This author speculates the strength of MER may appear in data gathered *in vivo* where additional signals may help models overcome the increase of noise in the data.

One way that this project aims to be a stepping stone is by evaluating emotional recognition using data that could be collected *in vivo.* A key limitation, however, is that, while the project aims to use this **type** of data, this project did not use data collected by common wearables or look at participants *in vivo*. A future experiment would involve receiving data from modern, commercially available smart devices, e.g. Apple Watch, and using that data along with self report to classify emotions. Using commercially available smart devices will also allow *in vivo* testing to determine if emotional recognition can be viable as people go about their daily lives.

4.4 Considered approaches for this project

Several approaches and methodologies were considered for this project that were ultimately not implemented due to concerns about preserving information in the data or feasibility of implementation.

One approach that was considered for this project was segmenting the data. It was initially considered to segment the data into smaller windows of time, e.g. 5 seconds, and extract features from each individual segment. This was decided against for this project due to concerns of losing information of HRV with the ECG signal. Segmenting the data into the R-R intervals could also be possible, though depending on size of data may lead to an impractical number of features. Segmentation, therefore, can be explored in future studies.

Another consideration of this project that was implemented, but ultimately decided against was dimensionality reduction using principal component analysis. This was implemented, however, accuracy did not improve significantly. It largely remained the same or decreased, and so it was ultimately decided against in favor of preserving information. It is possible that if changes were made to the data, such as the previously mentioned segmentation, then tools like principle component analysis may prove more helpful.

6. CONCLUSION

The growth of wearable technology opens up a wealth of new physiological data, collectible in real time, in everyday life, like never before. This project used ECG and GSR data from 12 children to develop emotional recognition models and evaluate early and late methods of data fusion. The key findings from this project are that early fusion performs better than a late fusion majority vote system by a large margin. Additionally, no single model outperforms the others, though the random forest classifier appears to have the strongest robust performance. Finally, models trained on fused data performed similarly to models trained on only a single physiological signal, and consequently multimodal methodologies are not inherently superior to unimodal methodologies.

With new data streams coming from the booming industry of wearable technology, there are many avenues to continue this research. Doing so may increase the accuracy of emotional recognition systems. Importantly it may facilitate the development of emotional recognition software that is adept at functioning in a normal everyday environment, allowing technology to become more affect aware. Such technology can aid humans in a number of ways such as facilitating positive outcomes in several key areas including education and healthcare.

7. REFERENCES

Darwin, C. (1872/1998). *The expression of the emotions in man and animals* (3rd ed.). New York: Oxford University Press.

Descartes. (1649/1988). The passions of the soul. In J. Cottingham, R. Stoothoof, & D. Murdoch (Eds.), *Selected philosophical writings of René Descartes* (Vol. 1, pp. 325–403). Cambridge: Cambridge University Press.

Ezzameli, K., & Mahersia, H. (2023). Emotion recognition from Unimodal to Multimodal Analysis: A Review. *Information Fusion*, *99*, 101847. https://doi.org/10.1016/j.inffus.2023.101847

Makowski, D., Pham, T., Lau, Z. J., Brammer, J. C., Lespinasse, F., Pham, H., Schölzel, C., & Chen, S. A. (2021). NeuroKit2: A Python toolbox for neurophysiological signal processing. *Behavior Research Methods*, 53(4), 1689-1696. https://doi.org/10.3758/s13428-020-01516-y

Malliani, Alberto, et al. "Cardiovascular neural regulation explored in the frequency domain." *Circulation* 84.2 (1991): 482-492.

M. N. Dar, A. Rahim, M. U. Akram, S. Gul Khawaja and A. Rahim, "YAAD: Young Adult's Affective Data Using Wearable ECG and GSR sensors," *2022 2nd International Conference on Digital Futures and Transformative Technologies (ICoDT2)*, Rawalpindi, Pakistan, 2022, pp. 1-7, doi: 10.1109/ICoDT255437.2022.9787465.

Pan, B., Hirota, K., Jia, Z., & Dai, Y. (2023). A review of multimodal emotion recognition from datasets, preprocessing, features, and Fusion Methods. *Neurocomputing*, *561*, 126866. https://doi.org/10.1016/j.neucom.2023.126866

Rainville P., Bechara A., Naqvi N., Damasio A.R. Basic emotions are associated with distinct patterns of cardiorespiratory activity. Int. J. Psychophysiol. 2006;61:5–18. doi: 10.1016/

Russell, James. (1980). A Circumplex Model of Affect. Journal of Personality and Social Psychology. 39. 1161-1178. 10.1037/h0077714.

8.APPENDICES

8.1 *Hyperparameters tested for each model using sk-learn GridSearchCV function*

Model	Parameter	What was tested (Range generating function where appropriate)
Multilayer Perceptron	Structure	(50, 2), (100,), (100, 2), (50,), (50, 20, 2)
	Alpha	np.arange(0.0001, 0.001, 0.0002)
	Activators	Tanh and Relu
	Solvers	Adam and Stochastic Gradient Descent
	Learning Rate	np.arange(0.2, 1, 0.1)
Random Forest	Tree Count	np.arange(10,105, 5)
	Leaf Size	np.arange(1,6)
	Bootstrapping	Both Bootstrapping and not bootstrapping were tested
Logistic Regression	Penalty	L2 and elasticnet
	Solver	Saga
	L1 ratio	0.15

Ridge Classifier	None	No special hyperparameters were tried with the ridge classifier
K Nearest Neighbors	Neighbor count	np.arange(2, 12, 2)
Support Vector Classifier	Kernel	Linear, poly, rbf, and sigmoid
	C	np.arange(0.5, 2, 0.5)
Adaboost Classifier	Number of Estimators	np.arange(10, 150, 10)
	Algorithm	SAMME

8.2 *Averaging model performance across single signal analysis and early fusion. Highest for each target is bolded.*

Prediction Target	Model	ECG	GSR	Fusion	Average Accuracy
Valence	MLP	68.42	77.63	77.63	74.56
	RF	**76.32**	69.74	**84.21**	**76.75666667**
	LR	68.42	76.32	65.79	70.17666667
	RR	68.42	**78.95**	68.42	71.93
	KNN	63.16	60.53	71.05	64.91333333
	SVC	69.74	**78.95**	69.74	72.81
	ADA	71.05	65.79	82.89	73.24333333
Arousal	MLP	69.74	65.79	**72.37**	**69.3**

	RF	68.42	**73.68**	61.84	67.98
	LR	64.47	65.79	61.84	64.03333333
	RR	67.11	63.16	60.53	63.6
	KNN	**71.05**	**73.68**	63.16	69.29666667
	SVC	60.53	68.42	53.95	60.96666667
	ADA	67.11	68.42	63.16	66.23
VA Axis	MLP	47.37	55.26	53.95	52.19333333
	RF	**50**	**60.53**	**65.79**	**58.77333333**
	LR	38.16	50	48.68	45.61333333
	RR	34.21	53.95	40.79	42.98333333
	KNN	46.05	53.95	42.11	47.37
	SVC	38.16	53.95	52.63	48.24666667
	ADA	43.42	52.63	51.32	49.12333333
Emotional Tag	MLP	47.37	47.37	50	48.24666667
	RF	**52.63**	50	51.32	**51.31666667**
	LR	44.74	43.42	51.32	46.49333333
	RR	44.74	47.37	50	47.37
	KNN	43.42	47.37	53.95	48.24666667
	SVC	51.32	**51.32**	51.32	51.32
	ADA	51.32	36.84	**56.58**	48.24666667