Emotion Recognition with physiological Fusion

Anthony Long

***Abstract*—With over 2 million total soldiers in the US Armed Forces, the government has an essential need to identify and treat pain with quick response times. Ideally, pain would be identified and treated immediately after it occurs, in real time. In this paper, a system is proposed to identify pain through use of decision-level fusion on physiological data. The results across two individual trials showed an accuracy of 73.33% for both males and females, which was higher than the individual scores for validation and testing data.**

I. INTRODUCTION

The United States Armed forces has a need to identify ways to identify pain of soldiers in real time, which can be accomplished via use of emotion recognition through use of physiological signals. In this paper, the aim was to successfully use majority voting in order to classify a person's binary pain state as 'Pain' or 'No Pain' by individually classifying with 4 metrics, and giving an overall classification based on the majority of the individual metrics. This work provides a way to determine a person's pain state which is critical for immediate real-time analysis and administration of medical aid.

The remainder of the paper shall be laid out as follows: First, some related works on identification of pain via alternative or similar methods will be discussed, including physiological and imaging/video methods classification methods. Next, methods will be discussed, including an explanation of the Random forest classifier that is utilized in the classification system. Then, experimental design and results will be explained and analyzed. Then a discussion and conclusion will be drawn. Finally, Future work will be identified.

Hu et al [1]. tested the feasibility of a mobile neuroimaging-based clinical framework, CLARAi, which utilized augmented reality (AR) and artificial intelligence (AI) for objective pain detection from the patient's brain in real time. they used a functional near-infrared spectroscopy to gauge subject cortical activity during acute clinical pain trials simulated by cold simulations on subjects with hypersensitive teeth. Their 3-layer neural network (NN) achieved an optimal classification accuracy at 80.37% for pain and no pain discrimination.

Khan et al [2]. explored a novel computer vision system that recognizes expressions of pain from videos by analyzing facial features. By extracting shape information using pyramid histogram of orientation gradients (PHOG) and appearance information using pyramid local binary pattern (PLBP) they were able to achieve a discriminative representation of the face. Their testing results were an improvement over state-of-the-art results, seen when using PHOG and PLBP combined with a 2-layer NN. The 2-layer NN while simpler computationally than that of state of the art, was able to obtain an accuracy rate greater than 96% on average.

Jang et al [3]. examined the differences of boredom, pain, and surprise, as well as how to recognize these emotions based on physiological signals. These emotions were induced through emotional stimuli, and measured through use of physiological signals, including electrocardiography (ECG), electrodermal activity (EDA), skin temperature (SKT), and photoplethysmography (PPG). Twenty-seven physiological features were extracted from the signals to classify the three emotions, and the discriminant function analysis (DFA) was used as a statistical method along with 5 machine learning algorithms to classify the emotions.
Their results showed a significant difference of key physiological responses across emotions, with the highest recognition accuracy of 84.7 % being obtained through use of DFA.

Campbell et al [4]. explored pain-based emotion classification by analyzing different time and frequency domain features derived from electromyogram (EMG), skin conductance (SCL) and electrocardiogram (ECG) readings taken from subjects in response to induced pain. Their work encompasses an exhaustive/interpretable feature selection protocol to allow for a generalizable feature set. Associations between features were visualized using a topologically informed chart of the physiological feature space, which was used to identify key sources of information that led to the formation of five main functional feature groups. Said groupings were used to extract further insight into observable autonomic responses to pain through a complementary statistical interaction analysis. Their work observed that EMG and SCL derived features could functionally replace those obtained from ECG.

Lopez-Martinez and Picard [5] presented a pain intensity measurement method based on physiological signals, used for recognition of nociceptive pain. By implementing a multi-task learning approach based on neural networks that accounts for individual differences in pain responses while still leveraging data from across the population, they were able to test their method in a dataset containing multi-modal physiological responses to nociceptive pain. Results showed that accounting for individual differences through MTL allowed for improved pain intensity recognition compared to other approaches even with limited usage of features.

Kessler et al [6]. presented a new modality for pain classification based on remote Photo-plethysmography (rPPG). Utilizing the rPPG for pain classification, Kessler et al evaluated the benefits of the three-color channels of the rPPG signal, which was filtered in multiple frequency ranges to extract the heart rate and the respiration rate as biophysiological signals. They then classified pain with Support Vector Machine (SVM) and Random Forest (RF) classifiers. Their resulting performance was compared to the electrocardiogram (ECG) and the respiration from the biosignal amplifier and facial landmark features from video,

and it was shown that the rPPG signal can be used for pain classification, in particular the lower frequencies.

Jiang et al [7]. developed a continuous pain monitoring method using multiple physiological parameters, including heart rate (HR), breath rate (BR), galvanic skin response (GSR) and facial surface electromyogram. They collected data from subjects under thermal and electrical pain stimuli. classified as no pain, mild pain or moderate/severe pain based on self-reporting. Artificial neural network classifiers were trained, validated and tested with the physiological parameters, with the average classification accuracy being 70.6%. The same method was applied to the medians of each class and accuracy and resulted in an improved accuracy of 83.3%. The facial electromyogram utilizing leave-one-out cross-validation (LOOCV) saw an improvement to 95.9%. GSR, HR and BR were in general better correlated to pain intensity variations than facial muscle activities.

Neiberg et al [8]. constructed an emotion-recognizing system ERMIS based on psychological studies of emotion and the nature of emotion in its interaction with attention. A neural network architecture was constructed to handle the fusion of different modalities including facial features, prosody and lexical content in speech. The artificial neural network ANNA was able to automatically classify emotional states driven by the multimodal feature input. ANNA's novel feature was "allowance of a feedback attentional loop designed to exploit the attention-grabbing effect of emotional stimuli to further enhance and clarify the salient components of the input stream". Results showed crucial differences between subjects as to the clues they pick up from one another pertaining to emotional states.

Mikuckas et al [9]. developed a HCI system for emotional state recognition with the goal of stressful state recognition by means of the heart rate variability (HRV) analysis, due to HRV being a non-invasive method of emotional recognition. The states they identified corresponded to real life scenarios such as a person sitting, walking, or changing their posture over time. They study the impact of the emotional state and posture impact on HRV. Time and frequency domain, as well as nonlinear parameters were calculated, and parameters most sensitive to emotional state were chosen. Variability of the HRV parameters were verified over time, with the results showing that posture has a great impact on the HRV parameters. Due to this, Mikuckas et al. integrated a posture detection subsystem in the HCI system. Their results showed A total classification accuracy rate of 71%, and they observed that a person's HRV parameters are subject to change over time.

Finally, Majid Mehmood et al [10]. Explored the possibility of emotional communication via brain-computer interface (BCI) systems for patients with neuropsychiatric disorders or disabilities. They identify emotional states by analyzing the features of electroencephalography (EEG) signals obtained from noninvasive EEG sensors that measure the electrical activity of neurons inside the brain and select the optimal combination of features for recognition. Using a 14-channel EEG machine, they gathered EEG data from subjects while they were shown images with four varied types of emotional stimuli (happy, calm, sad, or scared). Utilizing the Hjorth parameters (activity, mobility, and complexity) to measure signal activity of the time series data, they determined the optimal EEG features using a balanced one-way ANOVA. The features selected were able to outperform univariate and multivariate features. They further processed the optimal features for emotion classification using SVM, k-nearest neighbor (KNN), linear discriminant analysis (LDA), Naive Bayes (NB), RF, deep learning (DL), and four ensembles' methods (bagging, boosting, stacking, and voting). Their results showed that "the proposed method substantially improved the emotion recognition rate with respect to the commonly used spectral power band method".

II. Method

The methods involved include random forests, majority voting, and accuracy, precision, and recall metrics, as well as understanding a confusion matrix. Furthermore, downscaling and normalization are used, along with K-fold cross validation. Leo Breiman defines a random forest as "a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest" [11]. To understand this, it is first necessary to understand a decision tree. A decision tree is essentially a binary tree, where each node is a question to help classify the data. The answer to the binary question allows one of two paths to be taken to a sequential node, which asks another question. Following a decision tree to its depth allows the tree to classify the data in a binary manner, which works well for our task of classifying a person as either 'Pain' or 'No pain'.

A random forest then is simply a collection of decision trees, where from a pool of decisions, a subset is chosen randomly for each tree. Furthermore, a subset rather than the entire data is also sampled. The goal is to use all decisions in at least a single tree, while avoiding using all of them in one tree. This helps to prevent overfitting. However, the deeper a tree goes (i.e., the more decisions are asked) the more specific a tree becomes, and since binary choices go up at an exponential rate, often max tree depth is limited in order to prevent overfitting. By meeting these criteria, a random forest is able to have multiple decision trees takes random, uncorrelated subsets of the data and classify them based on the decisions each tree has. Different methods can be implemented in order to utilized all of the trees' individual classifications of the data, but often the decision that the majority of the trees have found is taken as the classification for a data subset.

Once a random forest has been used to classify individual data type (systolic BP, Diastolic BP, etc.) majority voting is implemented to determine the final classification of 'Pain' or 'No pain'. Majority voting is a type of decision level fusion, in that multiple decisions are fused together to make the final classification. For example, if 5 decisions are made for a yes/no question, and the decisions are 4 yes to 1 no, the majority decision of yes would be picked as the final classification.

The data in this experiment is also downscaled and normalized, which are both simple processes. Downscaling is a process by which a larger set of data is scaled to a smaller one. Once a downscaled size has been decided, a section of rows the size of the ratio is assigned to each member of the downscaled data, and the rows that correspond are averaged into the new row. In this experiment, the downscaling size is

5000 rows. So, consider a data subset with 10,000 rows. The ratio is 10000:5000, or 2:1. So for every one of the 5000 new rows, 2 of the original rows will be averaged together and assigned to the new rows. Normalization is simply taking the data values and assigning them numbers on a scale from 0 to 1. This is dependent on the highest and lowest values in the dataset. For example, if a data set ranged from -3 to 10, then the max value of 10 would be rescaled to 1, and -3 would be rescaled to 0. All values between would be appropriately scaled between 0 and 1. Additionally, the training data has K-Fold cross validation implemented on it, which is the process of splitting the data into subsets of training and validation sets. In this experiment, 3-Fold cross validation is used, meaning that for the 60 subjects in the data sets, 40 are used for training, and 20 are used for validation of the training data. The validation data is cycled to include all of the data, so for 60 3Fold with 60 samples, there are 3 runs of the data, where validation data is the first 20, the second 20 and the third 20, and the training for each of the runs is the remaining 40 samples.

Metrics used in the experiment include: **Accuracy**, a measure of the correctly classified data- these are the green cells identified in figures 3-5, when 'Pain' or 'No pain' are correctly classified as such out of all classifications; **Precision**, which is a measure of the correctly positively identified values out of all values- In the context of this study, the 'Pain' classified as 'Pain', out of the total number of positive predictions in the dataset; and finally, **recall**, which is a measure of the True positives ('Pain' classified as 'Pain') against the bottom column, which is the total number of actually positive samples in the data subset. The confusion matrix is simply a spread of the predictions generated by the classifier to help compute these metrics.

III. Experimental Design and Results

In The experimental design, 4 physiological metrics were tested, including diastolic pressure, systolic pressure, electrodermal activity (EDA), and respiration rate. Experimental design utilized data1.csv and data2.csv, which are two files containing physiological epoch timestamps of the 4 mentioned data types for females and males, respectively. Each participant had all 4 datatypes, and included two copies of each, one when the subject was in pain, and one when they were not, totaling for 8 rows of data per subject. There was a total of 30 subjects per file, resulting in a grand total of 240 rows of data per file.

The experimental procedure was executed twice, once with the male file as training data for 3-fold cross validation and the female file as the testing data, and once more with the files' roles reversed. Once training and testing was run, the classifications of each data types were compared with majority voting in order to give a classification that considered all data types in an attempt to be more accurate. For each users' 4 datatypes, a classification of 'Pain' or 'No Pain' was given, and of those 4, the classification that was given most for that user was assigned as their label. If There was a 2-2 split, then a random classification was assigned.

data1.csv as training, data2.csv as testing				
3-Fold	DIA	EDA	SYS	RES
Accuracy	70.00%	38.33%	75.00%	40.00%
Precision	70.90%	39.93%	78.63%	37.58%
Recall	66.67%	53.33%	73.33%	40.00%
Testing	DIA	EDA	SYS	RES
Accuracy	73.33%	66.67%	66.67%	43.33%
Precision	68.42%	62.50%	63.89%	42.86%
Recall	86.67%	83.33%	76.67%	40.00%

Figure 1. Accuracy, precision and recall scores for training and testing data with data1.csv as the training data, and data2.csv as the testing data.

data2.csv as training, data1.csv as testing				
3-Fold	DIA	EDA	SYS	RES
Accuracy	76.67%	61.67%	51.67%	63.33%
Precision	76.84%	64.27%	51.11%	65.16%
Recall	76.67%	53.33%	56.67%	60.00%
Testing	DIA	EDA	SYS	RES
Accuracy	70.00%	56.67%	73.33%	58.33%
Precision	77.27%	56.25%	79.17%	55.81%
Recall	56.67%	60.00%	63.33%	80.00%

Figure 2. Accuracy, precision and recall scores for training and testing data with data2.csv as the training data, and data1.csv as the testing data.

DIA_3-Fold	No Pain	Pain
No Pain	6.7	3.3
Pain	2.7	7.3

DIA_3-Fold	No Pain	Pain
No Pain	7.7	2.3
Pain	2.3	7.7

EDA_3-Fold	No Pain	Pain
No Pain	5.3	4.7
Pain	7.7	2.3

EDA_3-Fold	No Pain	Pain
No Pain	5.3	4.7
Pain	3	7

SYS_3-Fold	No Pain	Pain
No Pain	7.3	2.7
Pain	2.3	7.7

SYS_3-Fold	No Pain	Pain
No Pain	5.7	4.3
Pain	5.3	4.7

RES_3-Fold	No Pain	Pain
No Pain	4	6
Pain	6	4

RES_3-Fold	No Pain	Pain
No Pain	6	4
Pain	3.3	6.7

DIA_Testing	No Pain	Pain
No Pain	26	4
Pain	12	18

DIA_Testing	No Pain	Pain
No Pain	17	13
Pain	5	25

EDA_Testing	No Pain	Pain
No Pain	25	5
Pain	15	15

EDA_Testing	No Pain	Pain
No Pain	18	12
Pain	14	16

SYS_Testing	No Pain	Pain
No Pain	23	7
Pain	13	17

SYS_Testing	No Pain	Pain
No Pain	19	11
Pain	5	25

RES_Testing	No Pain	Pain
No Pain	12	18
Pain	16	14

RES_Testing	No Pain	Pain
No Pain	24	6
Pain	19	11

Figure 3. Confusion matrices for individual physiological metrics from 3-Fold validation data (left column) and testing data (right column).

Results from the experiment show that in general in the first trial when females were the training and males were the test data, that the classifier was able to classify with a higher accuracy, precision and recall. Conversely, when males were training data, the classifier had lower accuracy, precision and recall. While there are exceptions such as the

systolic results for testing in trial 2 being higher than those of the training, in general both figures 1 and 2 support these findings.

Majority	ALL				
Accuracy	73.33%			No Pain	Pain
Precision	69.44%		No Pain	25	5
Recall	83.33%		Pain	11	19

Figure 4. Accuracy, precision and recall scores for majority voting with data1.csv as the training data, and data2.csv as the testing data, as well as the confusion matrix.

Majority	ALL				
Accuracy	73.33%			No Pain	Pain
Precision	79.17%		No Pain	19	11
Recall	63.33%		Pain	5	25

Figure 5. Accuracy, precision and recall scores for majority voting with data1.csv as the training data, and data2.csv as the testing data, as well as the confusion matrix.

Majority voting results can be seen above from both trials, with precision, accuracy, recall and the confusion matrices being displayed. Based on the individual training and testing results, it can be seen that the majority voting scores are higher than the majority of individual scores, with exceptions in trial 1 including systolic training and diastolic testing accuracies diastolic and systolic training precision, and diastolic and EDA recall rates all being equal to or higher than the majority voting results. These values can be seen highlighted and green (above majority voting) and blue (below majority voting) in figures 1 and 2, compared against the accuracy, precision and recall values in figures 5 and 6. For trial 2, exceptions include the diastolic training and systolic testing accuracies, the systolic testing precision, and the diastolic training and systolic and respiration testing recall rates all being equal to or higher than the majority voting accuracy, precision and recall. Additionally, the confusion matrices of majority voting, as well as training and testing indicate the correct and incorrect classifications based on majority voting and testing/training respectively, determining whether a user can be classified as 'Pain' or 'No Pain'. Correct classifications are highlighted in figures 3-5 as green cells, while incorrect classifications are red cells.

IV. Discussion and Conclusion

Analysis of the results include multiple findings. It is of note that while on average the majority voting was more accurate than individual metrics, systolic results were able to classify with better accuracy, precision, and recall than majority voting when female physiological data was testing data. Furthermore, when male physiological data was used as test data, accuracy and recall was higher individually than those of majority voting accuracy and recall, with precision rate being within 1%. This is a clear indicator that while overall majority voting may be better than considering 1 random metric, there are particular physiological metrics that are more successful at classifying pain than others. Furthermore, the fact that the individual metric that exceeded majority voting was different in each trial could be an indicator that which metric proves the most successful is dependent on the biological gender of a person.

Another finding is that in some cases, training data rates exceed testing data, and in other cases the testing rates exceed the training. This could be due to a number of reasons, such as the training data being insufficient or the number of K-Folds not being optimal. The fact that the training data has a greater number of classifications with cross-validation could explain why in some cases, the accuracy gets better. Conversely, because majority voting algorithm selects randomly in the case of a tie of metric classifications, there is to some degree random chances for testing data to be more (or less) accurate than individual testing data. Lastly, these anomalies in data could come from the fact that two distinct groups are used to train and test. In each trial, one biological gender is used to train, and the other is used to test. Depending on biological norms, in general one gender's physiological 'tells' may be a subset of another. For example, if males are harder to classify via systolic pressure, then using males for the trainer will subject the classifier to a more rigorous threshold. Then, when using female data as testing data, the classifier would have an easier (or potentially harder) time classifying.

Based on the findings, it is clear that physiological data can be used for moderately successful pain recognition. However, more work must be done to identify which physiological metrics are best to identify pain, which metrics should be fused (if at all). Fusion is almost universally more successful than analyzing individual components, but in order for fusion to be highly successful, the individual metrics that are fused must have moderate success at classifying. Furthermore, more work must be done to determine what classifier is best for each biological gender and depending on the given data (if biological gender is known), different metrics must be used for fusion in order to conduct majority voting. Contextually, whether or not physiological data is a good metric for pain recognition is contingent on all of these factors.

V. Future Work

Physiological data has been proven to be usable to identify, a person's pain state in a binary classification system. However, this is still prone to issues depending on a wide variety of factors, including classifier training parameters, gender, and what physiological modalities are best for a particular individual. For future work, it is advised to address these issues. A theoretical future approach for pain recognition could also be possible through use of integrated BCI devices, such as Elon Musk's Neuralink. Such technology would not have the same issues as the ones inherent to our work's issues, as it would still be a general classifier, but it would be able to build its model based on data taken solely from the user in question. It would allow for dynamic classifier molding, that could adapt to which physiological modalities most accurately classifies a user's pain state. Furthermore, the user could provide feedback to their Neuralink to enhance and improve the classifier to better classify pain in the future. This system would give vast amounts of information to researchers pertaining to norms that most people have, allow for specialized classification models on an individual level without large amounts of time or effort on any companies' part, and most importantly would not reveal confidential medical information to a third party attempting to determine a person's pain state. The third party would only inquire as to a user's pain state, and the Neuralink would do the classification and respond to the third party, without the need to pass the data for analysis.

References

[1] Hu X, Nascimento T, Bender M, Hall T, Petty S, O'Malley S, Ellwood R, Kaciroti N, Maslowski E, DaSilva A Feasibility of a Real-Time Clinical Augmented Reality and Artificial Intelligence Framework for Pain Detection and Localization From the Brain *J Med Internet Res* 2019;21(6):e13594 URL: https://www.jmir.org/2019/6/e13594 doi: 10.2196/13594

[2] R. A. Khan, A. Meyer, H. Konik and S. Bouakaz, "Pain detection through shape and appearance features," *2013 IEEE International Conference on Multimedia and Expo (ICME)*, San Jose, CA, USA, 2013, pp. 1-6, doi: 10.1109/ICME.2013.6607608.

[3] Jang, EH., Park, BJ., Park, MS. et al. Analysis of physiological signals for recognition of boredom, pain, and surprise emotions. J Physiol Anthropol 34, 25 (2015). https://doi.org/10.1186/s40101-015-0063-5 Y. Kong, H. F. Posada-Quintero and K. H. Chon, "Pain Detection using a Smartphone in Real Time*," *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, Montreal, QC, Canada, 2020, pp. 4526-4529, doi: 10.1109/EMBC44109.2020.9176077.

[4] E. Campbell, A. Phinyomark, and E. Scheme, "Feature Extraction and Selection for Pain Recognition Using Peripheral Physiological Signals," Frontiers in Neuroscience, vol. 13, 2019. A. J. Kolber, "Pain Detection and the Privacy of Subjective Experience," *American Journal of Law & Medicine*, vol. 33, no. 2-3, pp. 433–456, 2007.

[5] D. Lopez-Martinez and R. Picard, "Multi-task neural networks for personalized pain recognition from physiological signals," *2017 Seventh International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, San Antonio, TX, USA, 2017, pp. 181-184, doi: 10.1109/ACIIW.2017.8272611.

[6] V. Kessler, P. Thiam, M. Amirian and F. Schwenker, "Pain recognition with camera photoplethysmography," *2017 Seventh International Conference on Image Processing Theory, Tools and Applications (IPTA)*, Montreal, QC, Canada, 2017, pp. 1-5, doi: 10.1109/IPTA.2017.8310110.

[7] M. Jiang, R. Mieronkoski, E. Syrjälä, A. Anzanpour, V. Terävä, A. M. Rahmani, S. Salanterä, R. Aantaa, N. Hagelberg, and P. Liljeberg, "Acute pain intensity monitoring with the classification of multiple physiological parameters," *Journal of Clinical Monitoring and Computing*, vol. 33, no. 3, pp. 493–507, 2018.

[8] D. Neiberg, K. Elenius, and S. Burger, "Emotion Recognition," *Computers in the Human Interaction Loop*, pp. 95–105, 2009.

[9] Mikuckas, A., Mikuckiene, I., Venckauskas, A., Kazanavicius, E., Lukas, R., & Plauska, I. (2014). Emotion Recognition in Human Computer Interaction Systems. *Elektronika Ir Elektrotechnika*, 20(10), 51-56. https://doi.org/10.5755/j01.eee.20.10.8878

[10] R. Majid Mehmood, R. Du and H. J. Lee, "Optimal Feature Selection and Deep Learning Ensembles Method for Emotion Recognition From Human Brain EEG Sensors," in *IEEE* Access, vol. 5, pp. 14797-14806, 2017, doi: 10.1109/ACCESS.2017.2724555.

[11] L. Breiman, Machine Learning, vol. 45, no. 3, pp. 261–277, 2001.