Physiological Pain Recognition

Kevin Girjanand

Abstract- The purpose of this experiment is to develop a pain recognition system that uses physiological data to identify pain within soldiers. Specifically, the system is meant to fuse the physiological data in order to determine if soldiers are in pain or are in no pain in real-time. By fusing the given physiological data into a reduced state, it allows for a more suitable assessment and precise classification accuracies of pain recognition. This experiment provides the possibility of developing an effective and efficient method of real-time pain recognition, a system that can and will prove extremely useful to medical examinations for patients in critical condition. The results for the various trials of this study yielded distinct values of classification for each data type scattered among the many subjects. The extent of this paper will examine the procedure and methodology under which this experiment was conducted, along with a discussion of the results obtained and how they correlate to the effectiveness of the physiological pain recognition.

I. Introduction

Emotional recognition has always been a system of complex construct, often due to need for intricate designs to harvest data and the desire for more accurate classifications. The use of physiological data as a method of classification has been a topic of increasing discussion for many years, however there is a lack of true systematic analysis regarding the process itself [5].

Quantitative analysis of pain is immensely difficult because it is such a subjective and individual sensation that can be caused by existing or potential tissue injury [3]. A soldier can sustain a minor muscle strain that is not observable from the naked eye or does not agonize them immediately, which can then evolve into a much more detrimental form of pain. So this bears the question, at what point does the level of pain become recognizable? Is it at the moment of diagnosis for pain or is it the original point at which the micro-injury took place? This dilemma leads into the focus of the experiment within this paper, as a specific method is utilized as the primary experimental format. For certain individuals or situations, using an effective method to quantify pain without self-report is key, so the focus is directed towards using physiological signals within the human body to evaluate pain [3]. The use of these physiological signals provide a newer and more systematic quantitative evaluation method, one that can have better recognition for the intensity of pain that can be induced on any individual. Some forms of physiological signals include heart rate, skin temperature, and respiration rate, all of which are stimulated by slight electrical pulses through three-channel biosensors in order to gauge their current state [3].

For the experiment conducted and described in this paper, there were four different physiological signals spread across the datasets, specifically *Diastolic BP*, *Systolic BP*, *EDA*, and *Respiration*. The fusion of this data allows for systemically compact process of classification, as most of the signal data can range in minute differences, making it obsolete to operate with excessive amounts of data. In comparison to other studies, the use of physiological signals weighs heavily into the medical field, particularly with a recent study on patients with sickle cell disease and the use of signals for classification [2]. In this case, pain can be mitigated thorough the use of various medications, however there is a lack of understanding as to how the dosage should be distributed, which leads back to the main purpose of this experiment. By utilizing machine learning, "both sequential and non-sequential probabilistic models can be developed and directed to appropriately infer levels of pain from these various physiological measures" [2]. The outputs of these models will provide a distinct and comprehensive outlook on the effectiveness and efficiency of physiological signals as a method of pain recognition.

II. Method

Random forests are a combination of tree predictors where each tree depends on the values of a random vector that is independently sampled, along with a similar distribution for all of the trees in the forest [1]. For the given dataset, after the data had been read into the network into their respective training and testing formats, random samples are selected. Each sample will have a decision tree constructed, which will then get a prediction result from each decision tree. A vote will be performed for each predicted result in order to gain a viable statistical probability, where there will then be a selection for the final prediction based on the prediction result with the most votes.

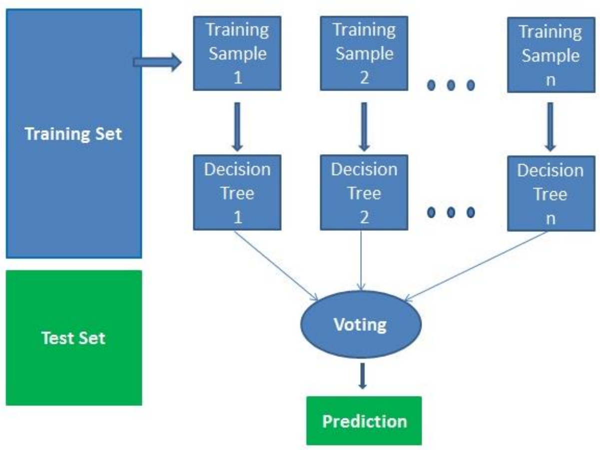

Figure 1: A visual representation of the random forest algorithm using data from a respective training and testing set.

As a method of statistical probability, random forests contain many advantages and disadvantages. Random forests are considered to be one of the more robust and highly accurate methods of classification as they require a large amount of decision trees for the process of classification. Due to this aspect, random forests eliminate the issue of overfitting, which is an error that occurs when a model becomes too complex to where it is unable to properly learn and analyze a dataset. Taking the average of all of the predictions cancels out any biases that may potentially occur, which makes this method very suitable for classification and regression problems. With all of the promising advantages provided by random forests, they can typically run very slowly with generating predictions, as there always exists multitude of decision trees. In addition, the model itself can be difficult to interpret in comparison to a decision tree, where you are able to easily decide on a choice by following a set path presented in the tree.

The construct of the classifier within this experiment relies on the fusion of data represented by the given physiological signals. "Multimodal information fusion seeks to greatly improve the performance of an inference model by smartly combining useful information that is extracted from a set of distinctive modalities" [4]. The performance and success of the designed architecture is contingent upon the ability to successfully integrate the extracted representations, however with each representation being specific to a single modality, finding the proper approach for effective information fusion can be time consuming and complex. Since pain recognition has gained newfound interest into the utilization of physiological signals, many studies and experiments rely on a set of carefully designed features that operate with common information fusion strategies. One study conducted extracted several distinct features from each input channel, mainly EDA, ECG, and EMG, to perform the necessary classification of varying levels of heat-induced pain, with a Random Forest classification model [4]. For the experiment detailed throughout this paper, a similar method of classifier is used to extract the most succinct features for the given dataset. This method will ultimately improve the ability to generalize the pain classifications from the different physiological signals as well as optimizing the efficiency of the model.

III. Experimental design and results

In the early designs of this system, the most crucial aspect was to firstly read in the physiological data to a logical format for analysis. This experiment was conducted with two distinct sets of data, both of which were used as the training set and the testing set for classification trials. Within each dataset, there were 30 individual subjects, each of whom had the four data types of *Diastolic BP*, *Systolic BP*, *EDA*, and *Respiration*, along with the two classes of pain and no pain. The physiological data itself was relatively substantial, often varying in length for each subject, which is not uncommon for this type of data. In order to properly apply a new system of fusion for the physiological data, certain procedures and methods were altered to make the data suitable for conducting the experiment.

For this specific system, the script was written in Python, a coding language that offers significant operations and functions to fulfill the necessary classification problems. To begin, the script will only run based upon two valid .csv files, both of which will serve as the training and testing sets for each run. Once entered by the user, a function *read_data* will be called twice for the respective datasets. The implementation of the function itself will take the given dataset as a parameter, opening the .csv file and begin reading it by row. Using the *reader* function from the *csv* library, the function will continuously iterate through all lines of the file, starting at the third index and running until the end of that specific line, appending a converted value from a string to a floating-point value to an empty list, *arr1*. That list will then be appended to another empty list, *arr2*, which will then reset the original "temporary" list *arr1* for future iterations. Another empty list, *arr3*, will append either a 0 or a 1 based on the given class type in the second index, "No Pain" or "Pain" respectively. The function will end up returning two separate lists, the first of which is actually a list of lists that contains the physiological data, and the other which holds a list of 0s and 1s to distinguish no pain or pain.

Another important function for the design of this system was the *normalize_data* function, which will conduct the down sampling and normalization of the data for necessary fusion. This function also uses a few Python libraries, namely *math* and *scikit-learn*, to produce the adequate outputs needed for classification. Using the list of physiological data values as the parameter, each list will be individually iterated, appending each value into a new empty list, *temp_arr*. From there, in order to down sample the data to an appropriate sample size, a *ratio* function was used to find a suitable reduction value. The length of each list was divided by the desired data length of 5000, using the floor value of that result as the ratio for the following resampling function. Once the down sampling occurred, the data was normalized using a *minmax_scale* function to get the values within a range of 0 to 1 within the list. That new list was then appended to another empty list, *new_arr*, and will continuously run until all of the physiological data has been adjusted and normalized, finally returning the "new" list of lists for data classification.

Now, for the sake experimentation, multiple runs were conducted using each data type, in which those results were saved for discussion and analysis. Each data type was trained and tested on a random forest, using the *RandomForestClassifier* function from the *scikit-learn* library, with a base setting of 100 estimators. This default value actually yielded the highest results within the closest range, as increasing the number of estimators caused the accuracy values to decrease significantly. On the contrary, decreasing the number of estimators actually yielded slightly higher accuracies for some of the data types, but greatly varied the range of the accuracies for others (One would be in the high 70s and another would be in the low 40s). Using this output from the *RandomForestClassifier*, another library function was utilized with that output, *VotingClassifier*, in order to use majority voting as a method of classification. Specifically, there was a single estimator from the *RandomForestClassifier* along with a setting of "hard" voting for majority rule voting. From there, the data was then fitted and predicted using the respective *fit* and *predict* functions, while also printing the accuracy of the score level fusion at the very end of the script.

Now for the results of this experiment, we will individually examine each data type and how they were classified. For the first data type, *Diastolic BP*, the classification accuracy for the first scenario (DS1 as training and DS2 as testing) ended with approximately 73% of correct classifications, with a precision of around 71% and recall of 80%.

```
Confusion Matrix (DS1 DS2):
[[20 10]
 [ 6 24]]
Accuracy: 0.7333333333333333
Precision: 0.7058823529411765
Recall: 0.8

Confusion Matrix (DS2 DS1):
[[20 10]
 [ 8 22]]
Accuracy: 0.7
Precision: 0.6875
Recall: 0.7333333333333333
```

Figure 2: Confusion matrices, accuracy, precision, and recall for BP Dia_mmHG

As for the other scenario (DS1 as testing and DS2 as training), the classification accuracy was exactly 70%, while the precision was lower at approximately 68.75% with a recall of 73%. *Diastolic BP* ended up producing the highest overall accuracies with the given datasets, seemingly remaining consistent with its overall values based on the outputted confusion matrices.

For the second data type, *EDA*, there was a more noticeable difference in the classification accuracies. For the first testing scenario, the classification accuracy was exactly 65%, with a precision of approximately 70% and a recall of 53%.

```
Confusion Matrix (DS1 DS2):
[[23  7]
 [14 16]]
Accuracy: 0.65
Precision: 0.6956521739130435
Recall: 0.5333333333333333

Confusion Matrix (DS2 DS1):
[[12 18]
 [11 19]]
Accuracy: 0.5166666666666667
Precision: 0.5135135135135135
Recall: 0.6333333333333333
```

Figure 3: Confusion matrices, accuracy, precision, and recall for EDA_microsiemens

The second testing scenario produced even lower values, with a classification accuracy of around 52%, a precision of 51% and a recall of 63%. The clear variance in the confusion matrices are a great visual representation of how the data is being interpreted, with a more accurate reading from DS1 as the training set as opposed to DS2.

For the third data type, *Systolic BP*, the results were more promising than *EDA* but still not ideal. The first testing scenario exhibited a classification accuracy of approximately 67%, with a precision of 72% and a recall of exactly 53%. The other testing scenario showed a slight increase in the classification accuracy, with it being around 72%, while having a precision of 68% and the highest recall overall, sitting at 83%. *Systolic BP* had a clear variance in its overall efficiency, as depicted in the confusion matrices in Figure 4, with the rows representing the predicted values and the columns representing the actual values.

```
Confusion Matrix (DS1 DS2):
[[24  6]
 [14 16]]
Accuracy: 0.6666666666666666
Precision: 0.7272727272727273
Recall: 0.5333333333333333

Confusion Matrix (DS2 DS1):
[[18 12]
 [ 5 25]]
Accuracy: 0.7166666666666667
Precision: 0.6756756756756757
Recall: 0.8333333333333334
```

Figure 4: Confusion matrices, accuracy, precision, and recall for LA Systolic BP_mmHg

With the first scenario, out of the 30 subjects for no pain, 24 were correctly classified by the model as no pain, however in the second scenario, 25 subjects with pain were correctly classified as pain. The intrigue with the datasets is that DS1 is more "favorable" towards no pain, with DS2 seemingly leaning towards pain for more accurate classifications.

For the fourth and final data type, *Respiration*, the results were the lowest overall. For the first testing scenario, the classification accuracy was approximately 67%, with a precision of around 71% and a recall of 57%. As for the second testing scenario, the classification accuracy and precision were both exactly 50%, with a recall of around 67%. *Respiration* has the lowest accuracy out of all the data types, only classifying about half of the subjects when DS2 is used as the training set and DS1 is used as the testing set.

```
Confusion Matrix (DS1 DS2):
[[23  7]
 [13 17]]
Accuracy: 0.6666666666666666
Precision: 0.7083333333333334
Recall: 0.5666666666666667

Confusion Matrix (DS2 DS1):
[[10 20]
 [10 20]]
Accuracy: 0.5
Precision: 0.5
Recall: 0.6666666666666666
```

Figure 5: Confusion matrices, accuracy, precision, and recall for Respiration Rate_BPM

After the experimental procedures were conducted with all of the data types, there were also distinct outputs for the results of each subject classification. A list was printed out for each data type of size 30, with a series of 0s and 1s to represent the classification of no pain and pain for a respective subject. These four lists were then integrated into a final list that used majority voting as a method of final classification. Any value had the majority of occurrences would be the overall classification, so a sequence of three 0s and a single 1 would be 0, and vice versa. For instances of equal occurrences where majority voting is not applicable, there is a choice of random selection for the final classification.

```
 DS1 DS2 Majority Voting Results (S represents 2 pain and 2 no pain)
[0 S 0 0 0 1 S 1 0 1 S 1 0 1 0 1 0 1 S 1 0 1 0 1 0 S 0 0 0 S 0 1 0 S S S 0
 1 0 S 0 1 S 1 1 0 0 0 0 0 S S 0 1 S 0 0 1 0 S]

 DS2 DS1 Majority Voting Results (S represents 2 pain and 2 no pain)
[0 1 1 S 0 1 0 S 0 S 1 S 0 S 1 1 1 1 S 1 0 1 1 0 S 1 0 S S S S S 0 1 S 0 1
 1 S 1 S 1 0 1 S 1 S S 0 0 S 1 S S 0 1 S 1 S 0]
```

Figure 6: Results from majority voting of testing and training splits

Figure 6 depicts the two lists for the testing scenarios that use majority voting as the method of classification, with 0s and 1s for no pain and pain, and an *S* used for simple visual distinction of values with no majority (This is mainly for understanding the effect of choosing no pain or pain). For both testing scenarios, the overall accuracy actually varied with multiple runs. The highest accuracy with majority voting went up to the lower 70s percentage-wise, while the lowest went down to the higher 50s percentage-wise. The average accuracy from majority voting was approximately 60% to 61%, which is not an ideal value for pain recognition on the subjects. There could be a multitude of reasons for such a low classification accuracy, one logical assumption being the system architecture. Potentially altering the number of estimators in the *VotingClassifier* function could potentially increase the accuracy of classifications, some of which were very low with the current construct. The data itself could also be difficult for the system to interpret, just based on the massive number of values that are stored within the dataset to represent physiological signals. So the systematic construct of this classifier model could in fact be improved with some tweaking of the functions used, but the current format does serve a sufficient purpose in the overall experiment.

IV. Discussion and conclusion

At the end of this experiment, it can be interpreted that using physiological signals as a method of pain recognition is a sufficient approach. Compared to classifying pain based on images, it is often very difficult to distinguish any level of emotion from visualization. But the internal recognition of what an individual feels should scientifically be more accurate, which leads back to the results obtained in this experiment. An interesting note is that the two data types that produced the highest classification accuracy were both of the blood pressures, *Diastolic BP* and *Systolic BP*, which may just be a coincidence for the naming, but could also be a sign of effectiveness. Reports and studies indicate that there is a correlation with an individual's blood pressure and chronic pain, so that does support the outcome of those two specific data types being the most accurate.

The aim for this experiment is to develop a more suitable classifier with machine learning to understand the physiological data in a more precise fashion. The fusion of data contributes greatly to the overall effectiveness and optimization of the system, as results are able to be systematically built upon combined data as opposed to using common, singular representations [5]. However, complications can arise during the implementation of data fusion, whether it can be a misrepresentation of data or lack of coherent fusion. In totality, fusion should only be considered and applied in situations where the data pool consists of a plethora of numerical values. Physiological data is one such manifestation, however even a logical application may not always derive the most optimal results, as the design and method of classification ultimately determines its true efficiency.

As for other approaches of pain recognition, there are not many alternatives that could viably outdo another. In general, pain is such an intricate representation that even in today's advanced state of science and research, there is no definitive method for pain recognition. Self-evaluations are the most common method to deduce whether a person feels pain, but the intensity may not be optimal for suitable interpretation. Even if it is, pain can range on a wide spectrum for every individual, making it robust to solidify a true method of assessment. However, a semblance of commonality can be observed from physiological signals as the primary method, holding it in a higher class than other forms of classification.

The approach of pain recognition using physiological signals was a modest choice of experimentation. Despite the obtained results not being optimal for true evaluation, there was still a determination of recognition with the given dataset. There can be improvements made to the classifier system in order to achieve higher classification accuracies, but the overall architecture operates in a sufficient manner that produces general results. The utilization of physiological signals has generated increasing interest within the field of emotional recognition, however there are still many more discoveries and advancements that need to be made before it can be acknowledged as a legitimate source of classification.

References

[1] L. Breiman, "Random Forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001

[2] M. J. Panaggio, D. M. Abrams, F. Yang, T. Banerjee, and N. R. Shah, "Can subjective pain be inferred from objective physiological data? Evidence from patients with sickle cell disease," *PLoS Comput. Biol.*, vol. 17, no. 3, p. e1008542, 2021.

[3] Y. Chu†, X. Zhao, J. Yao†, Y. Zhao, and Z. Wu, "Physiological signals based quantitative evaluation method of the pain," *IFAC proc. vol.*, vol. 47, no. 3, pp. 2981–2986, 2014.

[4] P. Bota, C. Wang, A. Fred, and H. Silva, "Emotion assessment using Feature Fusion and Decision Fusion classification based on physiological data: Are we there yet?," *Sensors (Basel)*, vol. 20, no. 17, p. 4723, 2020.

[5] P. Thiam, H. Kestler, and F. Schwenker, "Multimodal deep denoising convolutional autoencoders for pain intensity classification based on physiological signals," in *Proceedings of the 9th International Conference on Pattern Recognition Applications and Methods*, 2020.