Classification of Induced Pain Levels using ECG signals

Daniela Pais[1]
danielapais@ua.pt
Raquel Sebastião[1, 2]
raquel.sebatiao@ua.pt

[1]IEETA, DETI, LASI, Universidade de Aveiro, 3810-193 Aveiro, Portugal

[2]Polytechnic of Viseu, 3504-510 Viseu, Portugal

Abstract

Effective pain management depends on an accurate assessment of pain intensity. However, limitations in current pain assessment scales, including subjective reporting of pain and potential observational bias, can result in inadequate pain treatment. Therefore, in order to improve pain assessment and management, there is increasing interest in developing objective assessment methods, particularly employing physiological indicators.The aim of this work was to classify pain induced by a Cold Pressor Task (CPT) using features extracted from electrocardiogram (ECG) signals. The Random Forest algorithm demonstrated superior performance in distinguishing between low/moderate pain and high pain, employing a set of 15 ECG-features associated with the P, R, S, and T waves. This model achieved an accuracy of 95.3%, an F1-score of 94.0%, a precision of 97.9%, and a recall of 90.4%. These results demonstrate the feasibility of using physiological alterations in the ECG signal for assessing pain.

1 Introduction

An accurate assessment of pain intensity is crucial for effective pain management [4]. Currently, pain assessment scales rely heavily on self-reports from patients, and one widely used approach is the Numerical Rating Scale (NRS), where patients rate their pain from 0 to 10, representing the absence of pain to the worst imaginable pain [1]. However, these methods are subjective, discontinuous, and inadequate for evaluating the pain of patients unable or with limited ability to self-report [3].

Recent studies have provided evidence to support the use of physiological signals to develop strategies for objective pain assessment. The effects of the Autonomic Nervous System can be measured non-invasively through physiological signals, allowing for the detection of increased sympathetic activity related to pain through physiological changes rather than relying on self-report [2].

This study aimed to investigate alterations in electrocardiogram (ECG) signals induced by controlled pain elicitation through a thermal stimulation procedure known as the Cold Pressor Task (CPT), as a step towards developing an Artificial Intelligence (AI) system designed to provide objective pain assessment for supporting healthcare professionals in clinical settings. In this work, binary classification was performed between low/moderate pain (NRS score$<$8) and high pain (NRS score$\geq$8) levels, with the goal of identifying the most relevant ECG features and optimal models for accurately distinguishing these pain categories.

2 Methods

This section describes the experimental protocol employed for data collection and explains the methods implemented for analyzing ECG responses during cold stimulus-induced pain.

2.1 Dataset

The dataset comprises 642 examples and consists of data from 37 participants, 23 female and 14 male, with ages ranging from 19 to 25 years old (21.36 ± 1.27 years old).

2.2 Experimental Protocol for Data Collection

Initially, a five-minute baseline was recorded, and then participants were instructed to immerse their nondominant hand and forearm in a warm water tank for two minutes to ensure a consistent skin temperature across the participants before the CPT. After, the participants submerged their nondominant forearm in a cold water tank with a temperature of approximately $7^{\circ}C \pm 1^{\circ}C$. Participants were asked to endure the pain for as long as they could, with a time limit of two minutes. If they could not tolerate the pain, they were encouraged to inform the researcher and, before withdrawing their arm, to report their pain level. If they were able to complete the CPT, they were asked to report their maximum discomfort around the two-minute mark. Participants were required to report their pain level using the NRS. Afterward, participants were instructed to immerse their nondominant hand and forearm in the warm water tank for another two minutes. Before the end of the protocol, the participants were at rest for five minutes. The ECG was recorded continuously using minimally invasive equipment during the entire protocol. For further information concerning the experimental data collection procedure, including details on inclusion and exclusion criteria, ethical considerations, as well as the data collection setup, please refer to the publication cited in reference [5].

2.3 Methodology for Dataset Analysis

The experiments were performed in Python, mainly using scikit-learn.

2.3.1 Feature Extraction, Transformation, and Selection

The dataset includes 21 features extracted from the ECG signals (Table 1), computed based on the location of the peaks of the P, R, S, and T waves and the onsets and offsets from the P, R, and T waves (Figure 1). The features were extracted in 20-second periods with a 75% overlap and normalized by dividing each epoch by the average of the respective feature in the baseline. Furthermore, feature standardization was implemented for the classification models relying on distance measures.

Three learning settings were compared, including training the classification models with the set of 21 features, as well as with features selected through both filter and wrapper feature selection (FS) methods. The filter method is based on pairwise feature correlation, in which the feature with the lower variance was removed from each pair of highly correlated features. The wrapper method employed a backward elimination approach to sequentially generate feature sets ranging from 2 to 20 by iteratively removing one feature at a time from the original set of 21 available features. Although results were obtained for the feature sets ranging from 2 to 20, only the results of the learning setting that demonstrated the best performance for each classification algorithm will be presented in this work.

Table 1: Description of the extracted ECG features.

ECG Feature	Description
P,R,S,T_amplitude	Average amplitude of P ,R, S and T waves
P,R,S,T_distance	Average distance between each corresponding wave
P,R,S,T_peaks	Number of peaks of P, R, S and T waves
P,R,T_onsetamp	Amplitude of the onset of P, R, and T waves
P,R,T_offsetamp	Amplitude of the offset of P, R, and T waves
P,R,T_onoffdist	Average distance between the onset and offset of P, R, and T waves

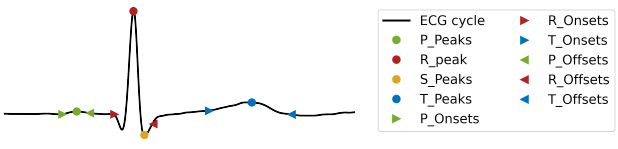

Figure 1: Location of the extracted peaks, onsets and offsets of the ECG.

2.3.2 Classification

To conduct binary classification, the samples were divided into two pain categories. Samples corresponding to high pain (NRS score$\geq$8) were assigned to the positive class, with 259 examples, while the negative class,

consisting of low/moderate pain levels (NRS score<8), included 383 examples. The dataset was divided into a training set with 513 samples (207 positive and 106 negative) and a test set with 129 samples (52 positive and 77 negative) in a stratified fashion considering an 80/20 split.

For comparing the classification algorithms, nested cross-validation (CV) was used on the train data. The test data was used for the final evaluation of the models. Six algorithms were evaluated, namely k-nearest neighbors (kNN), support vector machine (SVM), decision tree (DT), random forest (RF), adaptive boosting (AdaB), and extreme gradient boosting (XGB). The optimal hyperparameters of the algorithms were searched by maximizing the F1-score. In addition to the F1-score, accuracy, precision, and recall were also used to assess the generalization performance of each model.

2.3.3 Feature Importance

For the kNN and SVM algorithms, feature importance was assessed through feature permutation evaluation. Regarding DT, RF, AdaB, and XGB, feature importance was determined based on the total reduction of the criterion used for selecting the best split at each node.

3 Results and Discussion

The aim of this study was to analyze the physiological changes induced by pain in the ECG. Therefore, this study exclusively focuses on the data collected during the CPT, which corresponds to the phase of pain induction of the experimental protocol. The test results for pain classification are summarized in Figure 2. Overall, the models exhibited good performance in distinguishing higher pain from low/moderate pain, with SVM and DT models displaying the worst overall performance.

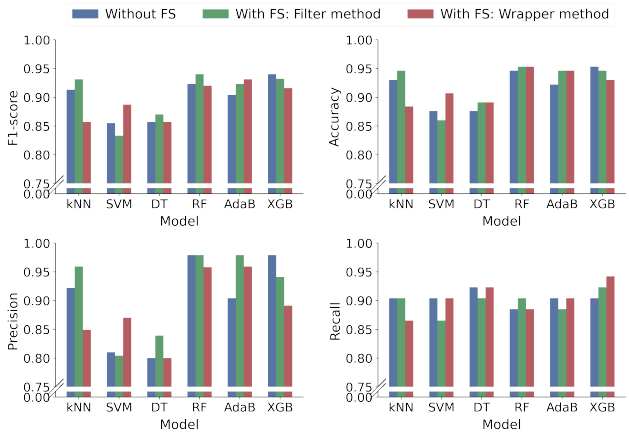

Figure 2: Performance of the classification models using the test dataset.

When training the models with the set of 21 features, the XGB model (learning_rate=0.5, n_estimators=500) demonstrated the best results, achieving an accuracy of 95.3% and an F1-score of 94.0%. kNN, SVM, AdaB and XGB models all achieved a recall of 90.4%, indicating that less than 10% of the high pain samples were misclassified as lower pain levels. XGB and RF models exhibited the highest precision (97.9%), indicating a high capability in classifying low/moderate pain.

The dataset resulting from FS based on pairwise feature correlation included 15 features (P,R,S,T_amplitude, P_distance, P_peaks, P,R,T_onsetamp, P,R,T_offsetamp, P,R,T_onoffdist). Employing these features, the RF model (criterion='entropy', max_depth=10, n_estimators=50) demonstrated the highest overall performance, with an accuracy of 95.3% and an F1-score of 94.0%. The six most significant features, sorted by importance, were P_offsetamp, T_onsetamp, P_onsetamp, R_onoffdist, R_offsetamp, and T_amplitude. While the RF model performs well in identifying high-pain samples, it may still incorrectly classify a small percentage of them as low/moderate pain, resulting in false negatives (FN) and the potential for missing a participant in pain (recall=90.3%). Both the RF and AdaB models showed a high precision of 97.9%, which indicates their ability to avoid false positives (FP).

Concerning the wrapper FS approach, the optimal number of features varied among the different classification models. While DT (20 features) and RF (18 features) required a larger set of features, kNN, SVM, AdaB, and XGB models performed well with a relatively smaller number of features ($\leq$ 15 features). The AdaB model (learning_rate=1, n_estimators=500) achieved an accuracy of 94.6% and an F1-score of 93.1%, using only 12 ECG features, with the most significant features being the amplitude (S,R,T_amplitude) and offset amplitude (P,T_offsetamp) of the ECG waves. Although the F1-score improvement compared to the previous approach, which employed 15 features, was only 0.8%, the reduction in the number of features not only reduces the model complexity but also results in faster run times. Furthermore, this model correctly identified 95.9% (FP=2) of the higher pain samples. The XGB model (learning_rate=0.1, max_depth=4, n_estimators=100) achieved the highest recall among all approaches, using 15 features, with a score of 94.2%, indicating that only 5.8% (3 samples) of the higher pain samples remained to be predicted. XGB attributed the highest importance to the onset amplitude of the T wave (T_onsetamp), the offset amplitude of the P wave (P_offsetamp), the amplitude of the T and P waves (T_amplitude, P_amplitude), the distance between corresponding S waves (S_distance), and the distance between the onset and offset of the R waves (R_onoffdist). Despite having an equal number of features (n=15) as the subset chosen through the filter method, the results obtained with this particular subset are inferior. Concerning the subset selected through pairwise correlation analysis, it yielded a precision of 94.1% (FP=3) and a recall of 92.3% (FN=4). In contrast, the subset obtained using SFS achieved a lower precision of 89.1% (FP=6) but demonstrated an improved recall of 94.2% (FN=3). This result underscores the importance of selecting the most relevant features for classification.

4 Conclusions and Further Research

Both undertreatment and overtreatment can result in psychological and physiological adverse effects. Thus, it is important to develop a pain management model that minimizes both FP and FN in assessing pain levels within clinical settings to ensure the effective management of pain. This study showed that ECG features related to the P, R, S, and T waves were effective in distinguishing between lower and higher pain. Nonetheless, the models exhibited superior performance in classifying lower pain samples, as evidenced by their higher precision scores compared to their recall scores. The RF algorithm, in combination with 15 ECG features, demonstrated the best overall predictive performance, with an accuracy of 95.3%, an F1-score of 94.0%, a precision of 97.9%, and a recall of 90.4%. The most significant features were P_offsetamp, T_onsetamp, P_onsetamp, R_onoffdist, R_offsetamp, and T_amplitude.

This work is an initial stage for an AI system that aims to support clinicians with an objective assessment of pain, which may also enable personalized healthcare. Future research includes investigating the ability of ECG signals for multi-class pain classification, exploring deep learning techniques, and considering the combination of various physiological signals for a multi-signal assessment to enhance the reliability and development of more effective pain assessment methods.

References

[1] Harald Breivik, P. C. Borchgrevink, Sara M. Allen, Leiv A. Rosseland, Luis Romundstad, E. K. Breivik Hals, Gunnvald Kvarstein, and Audun Stubhaug. Assessment of pain. *British Journal of Anaesthesia*, 101(1):17–24, 2008. doi: https://doi.org/10.1093/bja/aen103.

[2] R. Cowen, Maria K. Stasiowska, Helen Laycock, and Carsten Bantel. Assessing pain objectively: the use of physiological markers. *Anaesthesia*, 70(7):828—-847, 2015. doi: https://doi.org/10.1111/anae.13018.

[3] Pat Hummel and Monique van Dijk. Pain assessment: Current status and challenges. *Seminars in Fetal and Neonatal Medicine*, 11(4): 237–245, 2006.

[4] T. Ledowski, J. Bromilow, J. Wu, M. J. Paech, H. Storm, and S. A. Schug. The assessment of postoperative pain by monitoring skin conductance: results of a prospective study. *Anaesthesia*, 62(10):989–993, 2007.

[5] Pedro Silva and Raquel Sebastião. Using the electrocardiogram for pain classification under emotional contexts. *Sensors*, 23(3):1443, 2023. doi: https://doi.org/10.3390/s2303144.