Development of Machine Learning Algorithms Using EEG Data to Detect the Presence of Chronic Pain

Jonathan Miller, Skylar Jacobs, William Koppes, Frank Minella, Federica Porta, Fletcher A. White, M.S. Ph.D., Joseph A. Lovelace

Affiliations

Jonathan Miller, Skylar Jacobs, William Koppes, Frank Minella, Federica Porta, & Joseph A. Lovelace
-PainQx, Inc.
Fletcher A. White, M.S. Ph.D.
-Department of Anesthesia, Indiana University School of Medicine, Indianapolis, IN, USA.
-Stark Neurosciences Research Institute, Indiana University School of Medicine, Indianapolis, IN, USA.
- Research and Development, Richard L. Roudebush VA Medical Center, Indianapolis, IN 46202, USA.

Abstract: Chronic pain impacts more than one in five adults in the United States (US) and the costs associated with the condition amount to hundreds of billions of dollars annually. Despite the tremendous impact of chronic pain in the US and worldwide, the standard of care for diagnosis depends on subjective self-reporting of pain state, with no effective objective assessment procedure available. This study investigated the application of signal processing and machine learning to electroencephalography (EEG) data for the development of classification algorithms capable of differentiating subjects in pain from pain free subjects. In this study, nineteen (19) channels of EEG data were obtained from subjects in an eyes closed resting state, and ultimately data from 186 participants were used for algorithm development, including 35 healthy controls and 151 chronic pain patients. Signal processing was applied to identify noise free segments of EEG data and 6375 quantitative EEG (qEEG) measures were calculated for each subject. Various machine learning methodologies were applied to the data, with Elastic Net chosen as the optimal methodology. The final classifier developed using Elastic Net contained 34 qEEG features with non-zero weights. The classifier was able to differentiate pain versus no pain subjects with an accuracy of 79.6%, sensitivity of 82.2%, and specificity of 66.7%. The features used in the classifier were evaluated and found to align well with contemporary literature regarding changes in neurological characteristics associated with chronic pain.

Keywords: Electroencephalogram (EEG), chronic pain, prediction, machine learning.

1 Introduction

Using a chronic pain model introduced in the 2019 edition of National Health Interview Survey, approximately 50.2 million U.S. adults (20.5%) report pain on most or every day (Yong, 2022) with the most common pain locations being the back, hip, knee, and foot. Furthermore, respondents with chronic pain reported limitations in daily functioning and more workdays missed compared with those without chronic pain (Yong, 2022).

The International Association for the Study of Pain (IASP) defines chronic pain as pain that has persisted for more than 3 months and is associated with significant emotional distress and/or functional disability. In a 2010 study, the total yearly cost associated with the condition was found to be $560 billion to $635 billion in the US alone. That cost was composed of direct health care costs ($261 billion to $300 billion), days of work missed ($11.6 billion to $12.7 billion), hours of work missed ($95.2 billion to $96.5 billion), and lower wages ($190.6 billion to $226.3 billion) (Gaskin DJ, 2011). Considering these figures, the costs associated with chronic pain in 2010 were more than those associated with heart disease or cancer treatments.

The current standard of care for assessing chronic pain intensity is patient self-report using the Numeric Rating Scale (NRS) or the Visual Analog Scale (VAS). Subjective pain measurement may lead to a non-replicable, unreliable assessment of a patient's pain state, resulting in inappropriate treatment decisions and ultimately poorly managed pain. Furthermore, there are several patient populations that are unable to effectively communicate their pain via self-report, challenging clinicians to determine appropriate treatment paths (Herr, 2011).

There has been an abundance of research in recent years focused on identifying neurological signals associated with specific clinical conditions utilizing various modalities, including electroencephalography (EEG). As a modality for acquiring neurological data, EEG offers several benefits relative to other technologies such as Functional Magnetic Resonance Imaging (fMRI) or Positron Emission Tomography

(PET), including lower cost, non-invasiveness, and ease of use. Additionally, EEG, and in particular quantitative EEG (qEEG) data lends itself well to the use of artificial intelligence (AI) and machine learning (ML) for identifying neurological patterns and developing sophisticated models for classification of various nervous system disorders (Mussigmann,2022).

The purpose of this study was to develop technology to assess a chronic pain patient's current pain state objectively and empirically based on neurological signals obtained from EEG recordings. With support from a National Institutes on Drug Abuse (NIDA) Small Business Innovation Research (SBIR) grant, a multi-site research study was conducted, and algorithms were developed using qEEG data to objectively classify study subjects as "in pain" or "not in pain."

Following the development of chronic pain classification algorithms, the qEEG features selected for the algorithms using AI/ML were assessed and found to be congruent with contemporary domain knowledge and the published literature on qEEG and pain. Congruency with previous findings supports the notion there are qEEG features (neurological signatures) that are generalizable across chronic pain patients which can be used to characterize a patient's pain state.

2 Methods

2.1 Ethics

Approval was obtained from Advarra Institutional Review Board (study number Pro00042433) on March 12, 2020, and subsequently registered the study on www.clinicaltrials.gov (NCT04585451). Written informed consent was obtained from all participants prior to inclusion in the study.

2.2 Inclusion/Exclusion Criteria

Male and female participants ages 18-80 were included in the study, and pain subjects were required to meet the IASP definition of chronic pain. Subjects with neurological disorders or other conditions affecting neurological activity were excluded from the study. Additionally, patients who may have had a reason to misrepresent their pain (e.g., patients on workers compensation) were also excluded.

Study enrollment involved recruiting and screening potential participants from the Comprehensive Pain and Wellness Center (New York City, New York), Manhattan Restorative Health Sciences (New York City and New Hyde Park, New York), and Panorama Orthopedics & Spine Center (Golden, Colorado). The study was advertised at the various institutions via posted flyers and through solicitation by the study investigators (and their respective staff) at the institutions.

In total, 334 patients were recruited for the study who met eligibility criteria and provided informed consent. 308 participants fully completed the protocol including 54 healthy controls and 254 chronic pain patients. The chronic pain patient cohort included various indications as the primary diagnosis; 150 patients with back pain, 125 patients with joint pain, 87 patients with various other sources of pain, and 14 patients for whom a primary pain source was not indicated.

Of the 308 who were initially considered eligible for analysis, eight subjects were dropped due to non-conformity with expectations underlying the recruitment strategy. Three healthy controls reported being in pain at the time of the exam (not due to the exam itself) and five chronic pain subjects reported no pain

at the time of the exam. This represented a confound in the expected “in pain” versus “not in pain” truth vector associated with the control versus pain groups, and thus these eight subjects were not considered in the analysis.

Twenty-one subjects did not successfully complete EEG processing. This was largely due to anomalous electrode activity not detected at the time of recording. Electrode bridging, which causes multiple electrodes to read identical values, and single electrodes remaining at a fixed voltage for a short period of time (less than 2.5 seconds) were the primary causes.

2.3 Study Protocol

Clinical information relating to subjects’ pain history and functional impairment was collected using standardized instruments. Characterization of subjects’ mental state was performed using the PROMIS instruments for depression and anxiety.

For the study, EEG collection was done using the Zeto WR-19 wireless headset which has a dry electrode array with the specifications shown in Table 1. Electrode locations were in accordance with the international 10-20 system of electrode placement (Figure 1). Linked mastoids were used as reference.

Fifteen minutes of eyes closed resting state EEG data were collected. Collection began after the electrode array was placed on the subject’s head and all electrode impedances were verified to be within the manufacture defined acceptable range. Subjects were instructed to sit as still as possible and relax for 15 mins while the recording took place. Technicians monitored the subjects and intervened if the subject was moving excessively or began to enter a sleep state (i.e., exhibited head nodding).

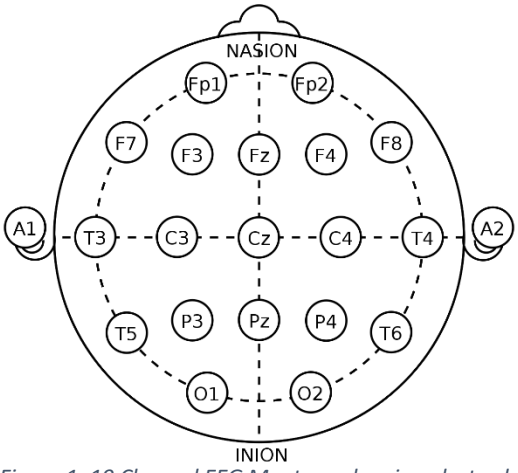


Figure 1: 19 Channel EEG Montage showing electrodes locations in accordance with the international 10-20 System.

Table 1. Summary of Zeto WR-19 Specifications

Specification	Value
Sampling Rate	500 samples/sec

Bandwidth	0.01 to 80 Hz
Data Resolution	24 bits
Noise Free Bits	19 bits

Table 1: Brainmaster Discovery 20 Specifications

2.4 EEG Data Processing

Following the 15 minutes of EEG data collection, the data was sent to a cloud computing platform, where it was processed by a series of software modules to refine the data to be used for classifier development (Figure 2). EEG data was supplied in a standard European Data Format (EDF) file. EDF is a standardized way of assembling the EEG signal information along with needed recording information such as sample rate.

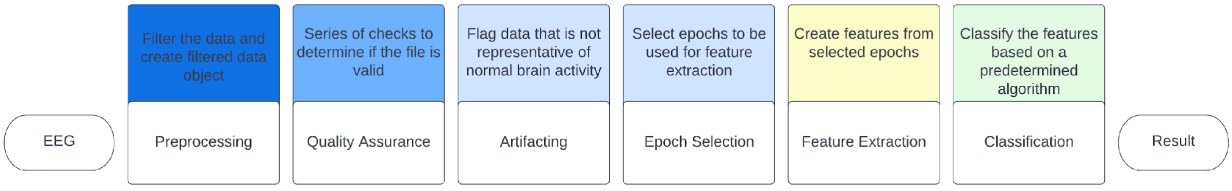


Figure 2: Processing Pipeline Diagram

2.4.1 Pre-processing

The first module in the pipeline filtered the raw EEG data by applying a second order Butterworth bandpass filter from 1-80 Hz. An Infinite Impulse Response notch filter was also used at 60 Hz to remove powerline noise.

2.4.2 EEG Data Quality Assurance

The second module in the pipeline assessed the quality of the EDF file containing the EEG data to ensure the file passed a series set of quality checks (Table 2). If any check failed, the data was deemed unusable due to a problem with the recording and a detailed report was created. The initial quality assurance check ensured the data files had a header and a body. The header was first checked to verify it was readable, consisted of data obtained from standard 10-20 electrode locations, and that frequency range, and dynamic range aligned with expected values. Subsequent checks operated on the body of the EDF which contained the time series EEG data. The data resolution (least significant bit), apparent sampling rate (derived from timestamps), and power spectral density were all checked against acceptable ranges. Individual electrodes were checked to ensure good contact quality over the course of the recording and variability was compared to all other electrodes. If more than 80% of the recording was deemed to be contaminated by artifact, the file was rejected. Only if all checks passed was the file advanced to the artifactor stage.

Table 2. EEG Quality Control Checks

Quality Check	Description of Check
EDF Header	Verify header is readable, standard 10-20 electrode locations are denoted, frequency range, and dynamic range align with signal data content
Recording Length	Recording length must exceed 12 minutes
Sampling Rate	Sampling rate must be 500 samples for second
Raw Dynamic Range	Raw dynamic range must be within the bounds of the amplifier
Minimum Step Size	Step between samples must be in line with the resolution of the amplifier
Percent Flat	Less than 10% of the recording can be flat, or showing no variability
Percent Clean	More than 20% of the recording must be deemed artifact free
Clean Dynamic Range	Once the recording is artifacted, dynamic range must be below ±300μV
Clean PSD vs TUH	Once the recording is artifacted, the recording is compared to a large corpus of publicly available EEG data and data must exist within the typical Power Spectral Density range from that corpus
Clean PSD vs Self	Once the recording is artifacted, each individual channel is compared to all other channels and all channels must be non-significantly different
Occipital Power	Once the recording is artifacted, the ratio of alpha power in the occipital vs the rest of the electrodes must be above 0.4.

Table 2: Table of Quality Control Checks

2.4.3 Epoch Creation

The full EEG recording was divided into segments referred to as “epochs,” each 2.5 seconds long. With a sample rate of 500 samples/second, each epoch contained 1250 samples across 19 electrodes. Successive epochs overlapped by 50%, so the EEG recording was effectively doubled in the array of epochs, since each epoch repeated half the information from the preceding epoch, and half from the following one. This approach offset the later use of a hamming windows as part of signal analysis: the window attenuated the signal down to zero at the beginning and end of each epoch, so the 50% overlap preserved signals attenuated near the start & end of each epoch. Various epoch lengths were tested during the algorithm development process, with the previously mentioned specifications resulting in the optimal performance during cross-validation.

2.4.4 Artifact Detection and Removal

The automatic artifactor software flagged any distortions in the EEG data caused by signals not originating from the brain. The artifactor produced a list of detected artifacts (Table 3), each of which was tagged with the onset time and duration in the recording. Each artifact caused an interval of EEG to be excluded.

Epochs were down selected for further processing based on artifactor results with any epoch overlapping in time with an artifact being excluded from further analysis.

Table 3. EEG Artifact Types

Artifact	Description
Impulse	Large excursion in amplitude, usually represented by a spike in the signal on one or more electrodes
Electromyographic	~30 Hz noise resulting from muscle activity on one or more electrodes
Vertical Eye Movement	Artifact resulting from the eye moving vertically because of blinking or eye flutter, localized to Fp1, Fp2, F7 and F8
Significantly Low Activity Signal	Signal is constant or near constant on one or more electrodes

Table 3: Table of Artifact types

2.4.5 Epoch Ranking and Selection

The epochs that remained after artifacting were then down-selected a second time based on covariance, a method derived from the work of Congedo, Barachant et. al (Congedo, 2017). For each epoch, the covariance matrix was generated, which is a 19 -by- 19 square diagonal positive-definite matrix (given the 19 electrodes used for data acquisition). The resulting array of matrices was then analyzed to determine a single covariance matrix representing the centroid covariance of all epochs, similar to finding the centroid of a cloud of points. All epochs with distance beyond the starting threshold of 6 were excluded. The process was then iterated: 1) find centroid of epochs under analysis, 2) calculate distance and remove epochs from analysis if their distance is beyond the threshold, 3) repeat. For each iteration, the threshold was reduced by 0.95. The intent was to determine the covariance cloud which represented the “normal” covariance for the recording. Iteration was stopped when a targeted percent of epochs remained in analysis, or when the remaining epoch cloud was tight (i.e., there were no more discarded epochs beyond the distance threshold). This resulted in a single centroid derived from the epochs remaining in analysis. Finally, distance was calculated to this centroid for all epochs, including those discarded from analysis, with the final calculation performed using Riemannian distance. The rational was that Euclidean distance was much less computationally complex, so it was used in the iterative process, while Riemannian distance was a superior choice for final evaluation, justified by the fact that covariance matrices are always positive definite, thus occupying a subspace of all possible matrices, and similarity between them is reflected by distance along the subspace manifold, calculated as Riemannian distance.

2.4.6 QEEG Feature Extraction

Epochs selected for further processing by the artifacting and epoch ranking modules were then passed to the feature extractor module. Over six thousand unique EEG features were calculated for each EEG recording. The "feature type" refers to the general type of mathematical analysis performed. Those analyses were carried out for all electrodes or electrode pairs as well as for each frequency band defined in table 4, which is what led to the large number of individual qEEG features shown in table 5.

Table 4. Frequency Band Descriptions

Frequency Band Name	Symbol	Frequency Range (Hz)
Theta	T	3.5 – 7.5
Low Alpha	A1	7.5 – 10.0
Alpha	A	7.5 – 12.5
High Alpha	A2	10.0 – 12.0
Beta	B	12.5 – 25.0
High Beta	B2	25.0 – 35.0
Gamma	G	35.0 – 50.0
Full Lower Spectrum	S	1.5 – 25.0

Table 4: Descriptions of frequency bands that were used in generation of features.

Table 5. QEEG Features Extracted

Name	Feature Type	Description	Number of Features
Alpha Power Ratio	Power	Ratio of power 9-11Hz divided by power 7-9Hz on all 19 electrodes, plus one additional alpha power ratio utilizing all electrodes. (Witjes, 2021)	20
Bipolar Power	Power	For each of the 7 primary bands, excluding S, power between all electrode pairs divided by total power across the S band. (John, 1990)	1197
Cross-Frequency Coupling	Connectivity	Similarity (correlation & coherence) between low frequency waveform components vs power envelopes of high frequency waveforms. (Canolty, 2010)	488
Coherence	Connectivity	For each of the 7 primary bands, excluding S, a measure of the statistical relation between all electrode pairs. (John 1990)	1368

Centroid Quartiles	Variability	For all 19 electrodes, variation in brain activity from moment to moment expressed as the distribution of distance from each epoch covariance to the centroid covariance of the full recording. (Barachant, 2013)	19
Granger Causality	Connectivity	Linkage between two electrodes in terms of the ability of one to improve forecasting the other. Only a subset of electrode pairs are utilized. (Marinazzo, 2011)	210
Mean Frequency	Frequency	The frequency at which half the power in the band is above and half is below within all 8 frequency bands for all electrodes and electrode pairs. (John 1990)	1520
Peak Dispersion	Frequency	For all 8 bands, the difference between mean frequency on non-occipital electrodes and the average of the mean frequency on occipital electrodes (O1 & O2). The average mean frequency of the occipital and the non-occipital electrodes are also included as features. (Halgren, 2018)	24
Power Asymmetry	Asymmetry	For each of the 8 primary bands and all 19 electrodes, ratio of power for all electrode pairs. (Wang, 2016)	1368
Monopolar Power	Power	For each of the 7 primary bands, excluding S, and all 19 electrodes the mean frequency divided by the S band. (John, 1990)	133
Spectral Event	Spectral Event	Excursions of brain activity in terms of their duration, bandwidth, peak amplitude, and frequency. Only a subset of electrode pairs are utilized. (Levitt, 2020)	28

Table 5: QEEG Features Calculated

2.4.7 Implications of Age on QEEG Features

Nearly all qEEG features have an expected trend with age. This was accounted for by converting features into Z-scores that incorporate age-expectation using a separate normative dataset from the Brain Research Laboratories of NYU School of Medicine which included 92 subjects with representation across ages 19 to 81. This norming dataset was used to fit a trend line for each feature. Subjects in this normative dataset did not have any evidence of clinically significant pain, depression, or anxiety, other than one individual with moderate anxiety. In each case, age dependence was assumed to be linear with log of age.

2.4.8 Classifier Development

The calculated qEEG features and participant reported NRS scores were used as input to AI/ML based classifier development methodologies with the objective of developing algorithms for the characterization of chronic pain. ML best practices were utilized in the development of a pain versus no-pain binary classification algorithm, to reduce the possibility of overtraining in which results are good on the training dataset, but poor in the broad population. The approach to mitigate overtraining included rigor in the cross validation so that test folds are truly isolated from the training folds, i.e. no information leakage was allowed to occur. In addition, ML results were compared to domain knowledge to identify solutions which do not align with relationships reported in literature, potentially representing something about the dataset that may be different from the broader population. Finally, 30% of the data was randomly set aside as an independent set of data to confirm the findings of the cross validation, and this 30% "hold out" group was only processed once at the end of development, producing a clean estimate of performance on the broad population.

2.5 Analysis of Classifier Performance

2.5.1 Train/Test and Hold Out Data Sets

Algorithms were developed on a Train/Test (TT) dataset (70% of the data), with performance evaluated on a Hold Out (HO) dataset (30% of the data) to check for potential overtraining of the algorithm. The performance of the Pain / No-pain classifier was analyzed using a variety of metrics. Performance was evaluated for both the TT and HO datasets. The TT performance was derived from 20 repetitions of 10-fold cross-validation, in which the TT set is partitioned at random into a Train interval and Test interval and results are aggregated, whereas the HO performance was a single application of the TT-developed classifier applied to the HO dataset.

2.5.2 Discriminant Score and Receiver Operating Characteristics Plots

The pain versus no-pain algorithm combined weighted qEEG values to produce a discriminant score. One method for visually assessing how well a binary classifier can separate two classes is a histogram of the discriminant scores for each of the two classes, which was generated as part of the study. Another method for visualizing the performance of a binary classifier, which was also utilized as part of the study, is the Receiver Operating Characteristics (ROC) curve. The ROC curve is a performance measurement for the classifier at all possible settings of the operating point, and so it represents an unbiased perspective on performance.

2.5.3 Area Under the Curve (AUC)

The Area Under the Curve (AUC) metric is derived from the ROC curve and represents the degree or measure of separability. It tells how much the model is capable of distinguishing between classes. The higher the AUC, the better the model is at predicting negative (no-pain) classes as negative and positive (pain) classes as positive. By analogy, the higher the AUC, the better the model is at distinguishing between patients with the disease/condition versus without the disease/condition.

2.5.4 Operating Point Selection

In order to generate specific performance metrics based on the AUC – ROC curve, an operating point must be established which determines the discriminant score cut point that defines which cases are classified

as positive (pain) versus negative (no-pain). The selection of an operating point is often made based on a risk/benefit analysis for a specific application of the classier, where false positive versus false negative results may have very different risks associated with them. Since a pain vs no-pain classifier may see several different applications, the study utilized an operating point that minimized the total number of false classification results.

2.5.5 Performance Metric Calculations Using Study Data

Having chosen an operating point, application of the classification algorithm yields a specific set of true positive (TP), true negative (TN), false positive (FP) and false negative (FN) results/counts. Those counts are often displayed in a 2x2 table referred to as a confusion matrix. Utilizing the data from the confusion matrix, a variety of performance measures were calculated, each providing a different perspective on the behavior of the algorithm. Definitions for the metrics calculated as part of the study include the following:

Accuracy: Percent of correctly classified cases involving both pain and no-pain classifications

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Sensitivity: Percent of correctly classified pain cases

$$Sensitivity = \frac{TP}{TP + FN}$$

Specificity: Percentage of correctly classified no-pain cases

$$Specificty = \frac{TN}{TN + FP}$$

Positive Predictive Value: Percent of subjects classified as in pain who are actually in pain

$$PPV = \frac{TP}{TP + FP}$$

Negative Predictive Value: Percent of subjects classified as no-pain who are not in pain

$$NPV = \frac{TN}{TN + FN}$$

2.5.6 Extended Performance Metric Calculations Using Study Data and Prevalence Models

Several of the performance metrics calculated, such as sensitivity and specificity, involve only one class, and so they are not impacted by the distribution of pain versus no-pain cases in the study population. Other metrics, such as Positive Predictive Value (PPV) and Negative Predictive Value (NPV), are impacted by changes in the ratio of positive (pain) to negative (no-pain) cases in the population under test. PPV and NPV often play an important role in the assessment of benefit versus risk associated with the use of a classifier for a specific purpose in a specific population. For this study, the goal of subject recruitment was to achieve a reasonable balance between classes (pain and no-pain) in order to optimize the ML process.

While that goal was achieved, the resulting study population was not representative of any of the populations that might be targeted for pain / no-pain assessment in the "real world." Therefore, in order to derive relevant PPV and NPV metrics, the Sensitivity and Specificity measures defined above were used to derive new confusion matrix values based on two "real world" prevalence models for pain versus no-pain. The first prevalence model was based on the incidence of chronic pain across adults in the U.S., which is reported to be approximately 20% (with 80% not in pain) (Yong, 2022). The second prevalence model targeted the population of patients being seen in U.S. pain clinics. That model was developed based on discussions with pain clinicians and assumed that 95% of patients being seen are experiencing pain (with 5% not in pain). Application of the sensitivity and specificity measures across each a given population resulted in a new confusion matrix consisting of derived values referred to as TP^x, TN^x, FP^x, and FN^x. Those new sets of counts were used to derive PPV and NPV numbers specific to the defined population.

Positive Predictive Value (for population x):

$$PPV(x) = \frac{TP^x}{TP^x + FP^x}$$

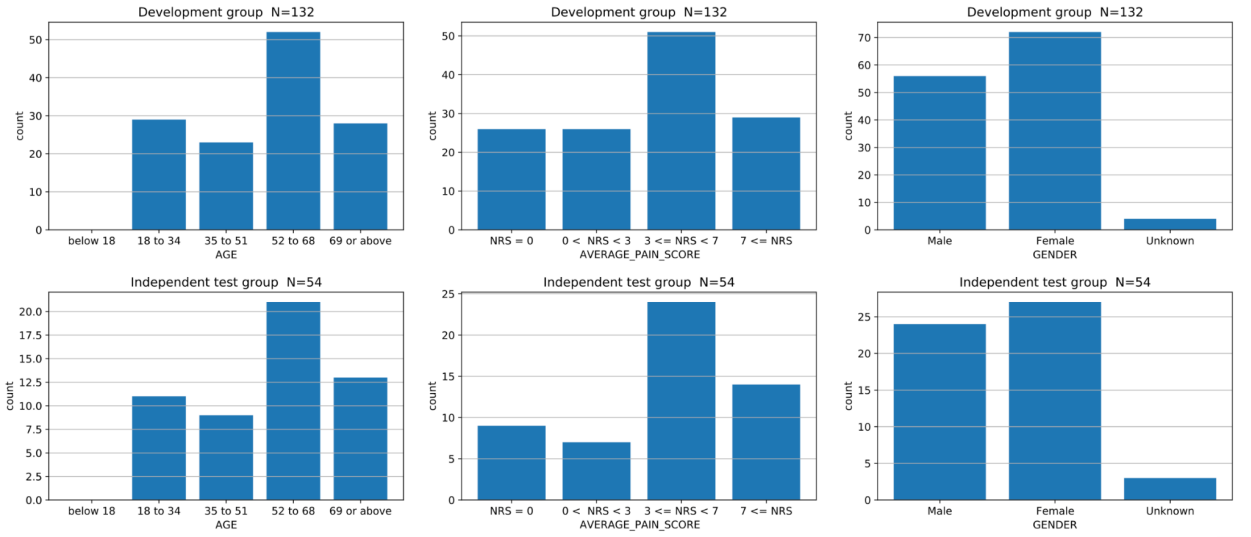
Negative Predictive Value (for population x):

$$NPV(x) = \frac{TN^x}{TN^x + FP^x}$$

3 Results

3.1 Data Set Generation Results

The data from 132 subjects (70% of the data) was used for algorithm development and the remaining 54 (30% of the data) was used as an independent test set. Participants were assigned to either group randomly, but stratified so that distributions of gender, age, and pain intensity were similar. (Figure 4)



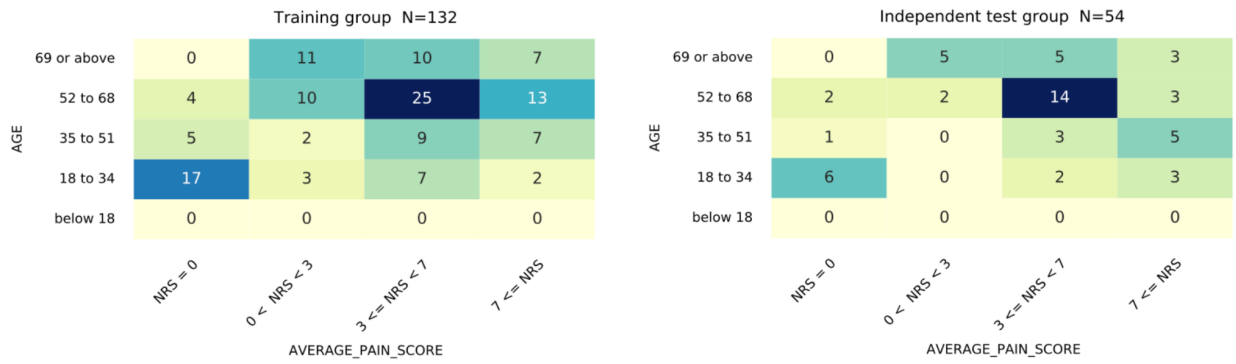


Figure 3. Histograms show the distribution of age, pain intensity, and gender. Heatmaps show joint distribution of age and pain intensity in the development (TT) dataset and the independent (HO) test dataset. Subjects were assigned randomly into these groups using stratification to ensure these distributions are similar. HO recordings were set aside and processed only after classifier development was completed using only TT.

3.2 Classification Algorithm Development Results

Given that the optimal ML methodology for a given classification problem can differ depending on the underlying phenomenon driving the differences between the defined classes, a variety of ML tools were applied including Gradient Boosted Trees, Nearest Neighbor Models, Logistic Regression Classifiers, Support Vector Machines, and Elastic Net Classifiers, each of which included many settings which were explored in detail. For example, Support Vector Machines were implemented with four different kernel types: Radial Basis Function, Polynomial, Sigmoid, and linear. From this work, we selected Elastic Net (ENET) (Hastie, 2009) as our preferred tool. Cross validation results were the dominant criteria for selection, but we also inspected details of the solutions to assess the contribution of different inputs (features) and whether highly weighted qEEG features met with expectation based on the relationships between neurological changes and chronic pain reported in the literature.

Elastic Net is a generalized method that encompasses two well-established methods: Ridge Regression and LASSO. Ridge Regression seeks to minimize the total magnitude of assigned weights (contributions) across all inputs (features), and LASSO seeks to minimize the total number of non-zero weights. ENET allows a smooth transition between the two so the objective function can reflect both goals in order to better fit a wide range of problems. The study utilized an L1_ratio of 0.5, meaning the solution was halfway between pure Ridge Regression and pure LASSO, and an Alpha parameter of 0.1, meaning 10% of the objective function was driven by driven by minimizing the number of features receiving non-zero weights, and the remaining 90% was driven by model match to data (i.e., least squared error).

3.3 Discriminant Score Results

Results for pain and no pain are shown in Figure 5. The left panel shows histograms of discriminant scores from the Development group (N=132), which arise from two hundred repetitions of 10-fold cross validation, and the right panel shows discriminant scores for the final classifier applied once to the independent test group (N=54).

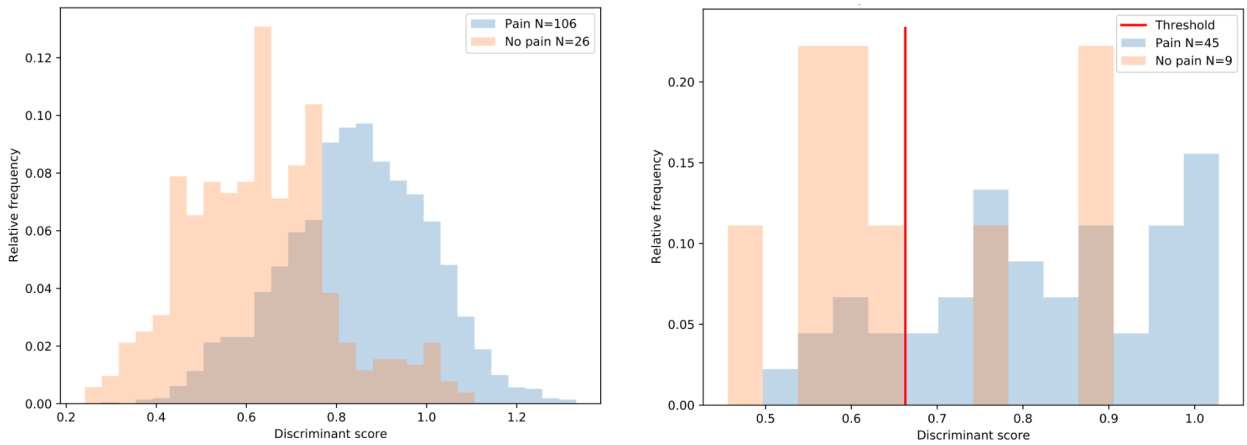


Figure 4. Histograms of discriminant scores for the development dataset (left panel) and the independent test set (right panel). The left panel shows test scores from the cross validation in which N=132 subjects each contributed 20 scores reflecting 20 repetitions of 10-fold cross validation. The right panel shows a single application of the classifier to an independent test set of N=54, which was processed only once.

3.4 Receiver Operating Characteristics (ROC) and Area Under the Curve (AUC) Results

The ROC curves for the TT and HO groups are shown in Figure 6. The ROC curves demonstrate the ability to classify pain versus no-pain subjects using QEEG features. The AUC calculated from the development group (TT) was 0.83, calculated as the average AUC in test subsets across 20 repetitions of 10-fold cross validation, and AUC in the independent test group (HO) was 0.78, calculated from the single application of the final classifier trained by the development group (TT).

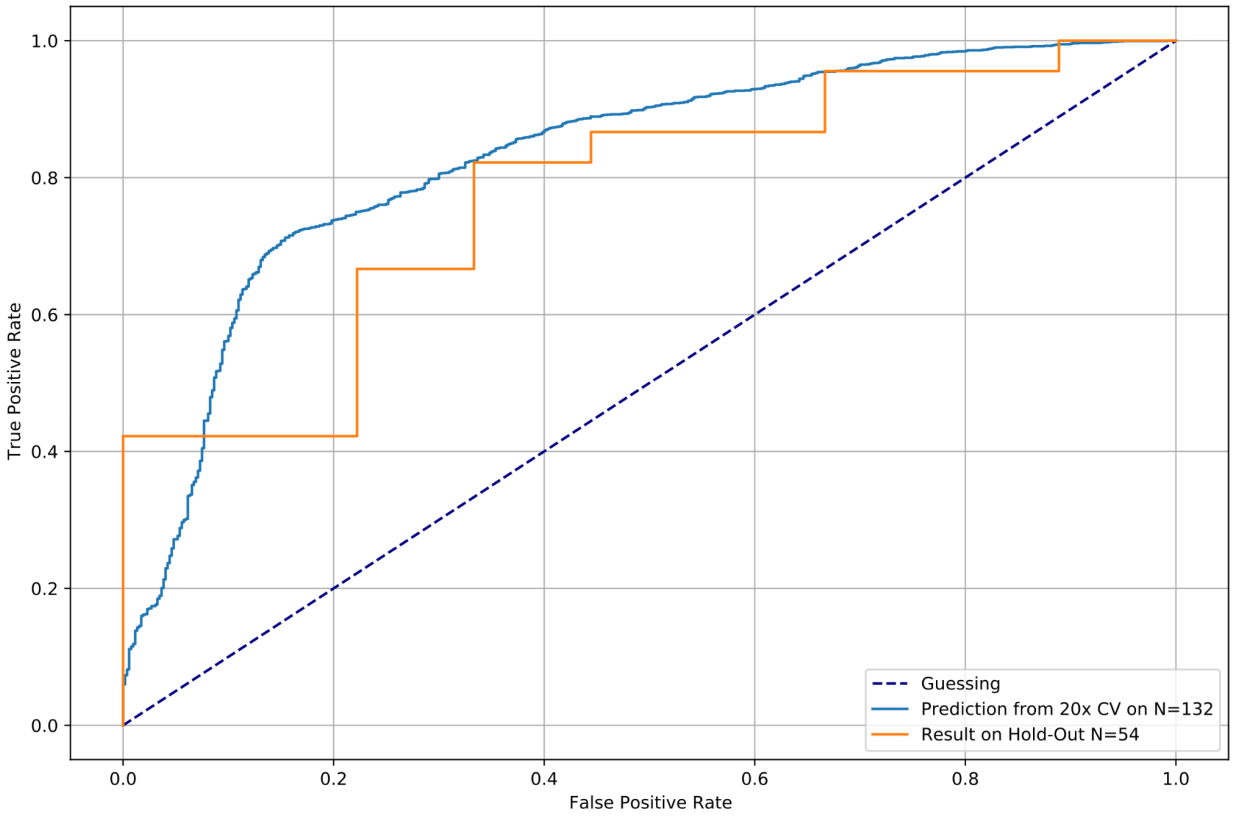


Figure 5. Receiver Operator Characteristic curves for the development dataset and the independent test set. The blue line is based on cross validation in which N=235 subjects each contributed 20 scores reflecting 20 repetitions of 10-fold cross validation. The orange line shows a single application of the classifier to an independent test set of N=151, which was processed only once.

3.5 Confusion Matrix Results

With the operating point set to minimize total false classification results, confusion matrix results for both the TT and HO groups are illustrated in figure 7.

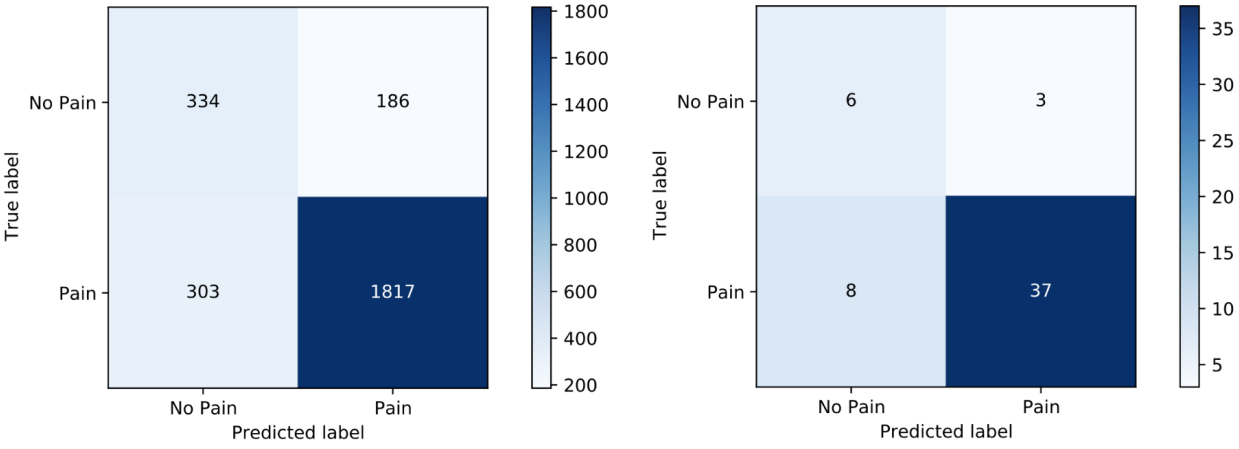


Figure 6. Confusion matrices show subject counts in the development dataset (left panel) and independent test (right panel). The left panel shows discriminant scores for subjects in the test interval of the cross validation, so each subject contributes 20 values to this matrix since the CV was repeated 20 times. The right panel shows subject counts for the single application to the independent test set.

3.6 Classification Algorithm Quality Metrics Results

Processing the data from the confusion matrices as described in the methods section of this paper, accuracy was 0.818 for TT and 0.775 for HO. Sensitivity/specificity was 0.858/0.654 for TT and 0.822/0.667 for HO. The full set of quality metrics for both the Train/Test and Holdout groups are shown in Tables 6. Confidence intervals shown in Table 6 reflect the variability in the twenty repetitions of 10-fold cross validation.

Table 6. Pain / No-pain Classifier Quality Metrics

Metric	Train / Test Dataset (TT)	Holdout Dataset (HO)
AUC	0.829 ± 0.016	0.775
Accuracy	0.815 ± 0.022	0.796
Sensitivity	0.857 ± 0.022	0.822
Specificity	0.642 ± 0.050	0.667
NPV	0.194 ± 0.029	0.165
PPV	0.978 ± 0.003	0.979

Table 6. Foundational Quality Metrics

3.7 Extended Quality Metric Utilizing Prevalence Models Results

The study population included 151 pain cases and 35 no pain cases. Thus, the prevalence of pain in the study population was 81.2%. In order to estimate PPV and NPV for populations in which the pain versus no-pain classifier might be utilized, two prevalence models were applied as described in the methods section. Table 7 summarizes the results when the two prevalence models were applied to re-calculate PPV and NPV.

Table 7. PPV & NPV Adjusted for Chronic Pain Population

Metric	General US Adult Population	Outpatient Pain Clinic Population
PPV	0.37	0.99
NPV	0.95	0.06

Table 7. PPV and NPV adjusted based of the prevalence of chronic pain in the general US adult population (20% prevalence) and a representative pain clinic population (95% prevalence)

3.8 QEEG Feature Selection and Weighting Results

Using ENET, specific qEEG features were selected and assigned a weight for use in the pain versus no-pain classification algorithm, with the weight assigned to a feature indicating that feature's relevance to the classification of Pain vs No-pain. Table 8 illustrates the top 10 most highly weighted qEEG features.

Table 8: Top 10 Features Ranked by Weight

Features rank	Feature	Frequency Band	Band Range	Electrodes		Weight	Contribution to Score
				#1	#2		
1	Bipolar relative power	Theta band	3.5 - 7.5 Hz	F4	F7	-0.0520	11.5%
2	Mean Frequency	Alpha band	7.5 - 12.5 Hz	C3	Cz	-0.0377	8.4%
3	Mean Frequency	Alpha band	7.5 - 12.5 Hz	F3	Fz	-0.0343	7.6%
4	Asymmetry	Beta band	12.5 - 25 Hz	F7	T5	0.0249	5.5%
5	Asymmetry	Alpha band	7.5 - 12.5 Hz	O1	O2	0.0247	5.5%
6	Asymmetry	High Beta band	25 - 35 Hz	P3	O1	0.0240	5.3%
7	Coherence	High Beta band	25 - 35 Hz	C3	T3	0.0233	5.2%
8	Asymmetry	Beta band	12.5 - 25 Hz	O2	F7	-0.0230	5.1%
9	Coherence	Low Alpha band	7.5 - 10 Hz	T3	T6	0.0217	4.8%
10	Asymmetry	High Alpha band	10 - 12.5 Hz	P3	Pz	0.0199	4.4%

Table 8: Top 10 features ranked by weight.

4 Discussion

4.1 Performance Metrics

The overall performance of the pain versus no-pain classifier developed through the application of ML on qEEG data indicates that EEG can serve as a foundational component of an algorithm for the objective determination of a chronic pain patient's pain state. For both the Train/Test and Hold Out datasets, histograms of discriminant scores illustrated differentiation of the pain versus no-pain groups. The ability to effectively separate the two groups was further supported by the ROC curves and AUCs for the Train/Test and Hold Out datasets, with an AUC of 0.83 for TT and 0.78 for HO.

The separation of the pain versus no-pain groups demonstrated by the discriminant score histograms and ROC curves enabled classification accuracy of 0.815 for TT and 0.796 for HO. Sensitivity/specificity was 0.857/0.642 for TT and 0.822/0.667 for HO. While potential risks associated with false negative and false

positive results would need to be mitigated for any commercial application of the classifier, these results are promising.

When mapped to a representative population of pain clinic patients, the sensitivity and specificity results achieved led to a Positive Predictive Value (PPV) of 0.979. This implies that if a pain clinic has any concerns regarding objective assessment of a pain patients pain state, a positive result could be relied on to identify patients in pain. On the other hand, the Negative Predictive Value (NPV) for the same population was 0.194, indicating that the classifier should not be used to identify patients who are not in pain. However, a negative (no-pain) result could be used to identify patients that require further assessment of their pain state. Since that assessment would probably be the same process currently applied for all pain clinic patients, an overall improvement in the efficiency of patient evaluation could potentially be achieved.

4.3 QEEG Features

Of the 6375 features calculated and submitted to Elastic Net, 34 received non-zero weight. The rest were effectively discarded from the solution. Elastic Net aims to minimize the L1 norm of the weight vector (i.e., the count of non-zeros) which acts to make the solution sparse, and that helps not only to simplify interpretation and greatly reduce the burden of feature computation, but it also imposes regularization on the solution, which mitigates the risk of overtraining (Zou, 2005). The summed magnitude of weights assigned to the top 10 features accounted for 63% of the total, leaving only 37% for the remaining 24 features. For this reason, the discussion below focuses only on the 10 features which contributed the most to classifier output (Table 8).

We assessed alignment of the features selected to various findings reported in the literature by inspecting the weight assigned to each feature and checking whether its sign (+/-) agreed with reports on similar measures investigated by other researchers and published in peer reviewed journal articles. Our ML solutions were developed solely from data acquired as part of this study, and domain knowledge is entirely independent, so alignment between these two lines of evidence is a strong indication that the pain versus no-pain classifier is capturing fundamental physiology and is not overtrained on spurious signals in our dataset. This was also confirmed by results obtained on the independent Hold-Out (HO) dataset, which played no role in development, was processed only once, and demonstrated performance (AUC 0.78) which is comparable to prediction from cross-validation (AUC 0.83) on the development (TT) dataset.

In terms of the magnitude of weight applied to the features, the second and third heaviest were Mean Frequency features, which represent the frequency at which the total power within the specified band is split in half: total power above equals total power below. This is the center of gravity formula for mean frequency (Klimesch, 1993), which we used in this study. These are the only two Mean Frequency features in the top 10, and they're both in the Alpha band (8.5 - 12.5 Hz) with negative weight. This means a decrease in the measured value of these features acts to increase the discriminant score, which makes a positive pain assessment more likely.This aligns well with the consensus reached by independent research worldwide over the past two decades that neural oscillations in the Alpha band tend to be slower among chronic pain subjects, as reviewed recently by [Musigman, Pinhero].

There were two Coherence features in the top 10, each one representing coherence of waveforms from an electrode pair, and this signal relates to connectivity between brain activity in different cortical regions. Both these features had positive weight in the classifier, which means an increase in measured value acts

to increase the discriminant score, which makes a positive pain assessment more likely. As with mean frequency, these Coherence features align well with the consensus among pain researchers that connectivity tends to be higher in chronic pain subjects.

Five features in the top 10 represented asymmetry, which reflects spatial variability of brain activity, and there is no consensus on this kind of signal relative to chronic pain. Of the five, most have positive weight (4 out of 5), which means an increase in spatial variability in neural activity acts, mostly, to increase the likelihood of a positive pain assessment.

One feature in the top 10 was a power feature in the Theta band, calculated as Bipolar Relative Power on the electrode pair F4 and F7. This feature has a negative weight in the classifier. While contemporary literature indicates that Theta oscillations tend to increase in chronic pain subjects, this is a measure of relative power which does not respond to broad changes across all electrodes and bands, but instead only indicates changes in the given electrode pair relative to all others, which is not a focus of contemporary literature.

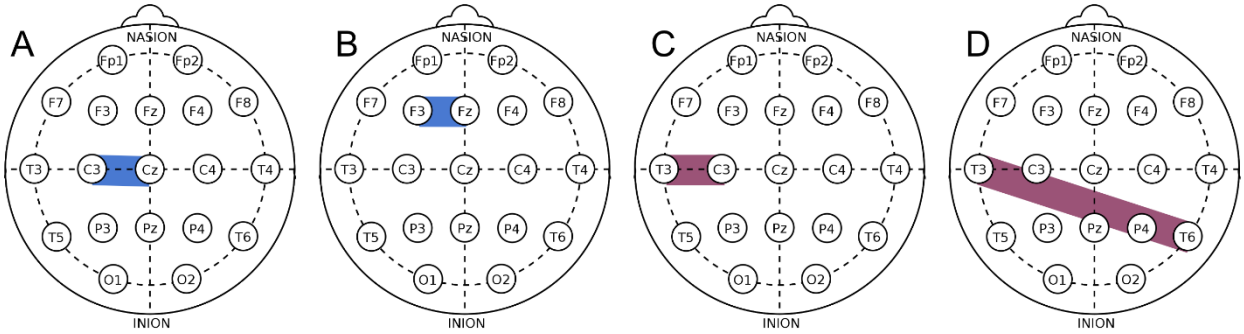


Figure 7: Out of the top 10 highest weighted features, two are based in Alpha band (8.5 - 12.5 Hz) shown in blue: A) Peak Alpha between C3 and Cz. B) Peak Alpha between F3 and Fz. Two others represent connectivity in different bands, shown in purple: C) Coherence between T3 and C3. D) Coherence between T3 and T6. Weights derived by machine learning show alignment with consensus in literature: Peak Alpha tends to decrease, and connectivity tends to increase with chronic pain.

5 Conclusions

This study successfully demonstrated that ML based algorithms utilizing EEG data can be used to differentiate subjects who are not in pain from subjects who are in pain. EEG features identified using the ML pipeline were congruent with domain knowledge identified by previous chronic pain and EEG studies. The ability to empirically differentiate subjects in pain from those not in pain has several potential applications, including determination of proper opioid dispensing practices and validation of chronic pain state in legal cases.

Even with the success of this research, limitations in the performance achieved indicate that additional research is needed to refine the technology. Although many qEEG features which were selected in the final algorithm align with domain knowledge, some features that have been reported to be relevant to chronic pain assessment were not selected. Additionally, the inherent subjectivity of the patient self-report and the natural variability of the human brain will always pose a limit to performance, but better understanding of the underlying mechanisms and neurological patterns related to chronic pain will allow us to approach or meet this limit. Areas of future research opportunities include conducting additional

studies to increase the size and diversity of the training database and combining EEG with other physiological data sources (such as heart rate or blood-based biomarkers) to further improve the technology and its pain intensity prediction capabilities.

6 Acknowledgements

Research reported in this publication was supported by the National Institute On Drug Abuse of the National Institutes of Health under Award Number R44DA046964. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

The author(s) declared the following potential conflicts of interest with respect to the research, authorship, and/or publication of this article: Frank Minella is the owner of PainQx and is an investor on patents licensed by PainQx from NYU School of Medicine. Jonathan Miller, Skylar Jacobs, William Koppes, Federica Porta, and Joseph A. Lovelace were employed by PainQx at the time of the research. Fletcher White consulted on the project and was an editor of the paper.

References

Acharya UR, S. V. (2015). A Novel Depression Diagnosis Index Using Nonlinear Features in EEG Signals. *European Nurology*, 79-83.

Barachant, A., Bonnet, S., Congedo, M., Jutten, C., 2013. Classification of covariance matrices using a Riemannian-based kernel for BCI applications. Neurocomputing 112, 172–178.

Canolty, R. T. (2010). The functional role of cross-frequency coupling. *Trends in cognitive sciences*, 506-515.

Cella, D. R. (2010). The Patient-Reported Outcomes Measurement Information System (PROMIS) developed and tested its first wave of adult self-reported health outcome item banks: 2005–2008. *Journal of clinical epidemiology*, 1179-1194. Retrieved from Healthmeasures.net .

Congedo, M. B. (2017). Riemannian geometry for EEG-based brain-computer interfaces; a primer and a review. *Brain-Computer Interfaces*, 155-174.

Fernando Soares de Aguiar Neto, J. L. (2019). Depression biomarkers using non-invasive EEG: A review. *Neuroscience & Biobehavioral Reviews*, 83-93.

Gaskin DJ, R. P. (2011). The Economic Costs of Pain in the United States. In *Institute of Medicine (US) Committee on Advancing Pain Research, Care, and Education. Relieving Pain in America: A Blueprint for Transforming Prevention, Care, Education, and Research.* (p. Apendix C). Washington DC: National Academies Press.

Halgren et. al. (2018). The generation and propagation of the human alpha rhythm. Proceedings of the National Academy of Sciences (PNAS), Vol. 116 No. 47.

Hastie, T. T. (2009). *The elements of statistical learning: data mining, inference, and prediction.* New York: Springer.

Herr, K. C. (2011). Herr, K., Coyne, P. J., McCaffery, M., Manworren, R., & Merkel, S. (2011). Pain Assessment in the Patient Unable to Self-Report: Position Statement with Clinical Practice Recommendations. Pain Management Nursing, 12(4), 230–250. doi:10.1016/j.pmn.2011.10. *Pain*, 230-250.

John, E. R. (1990). Normative Data Banks and Neurometrics. Basic Concepts, Methods and Results of Norm Constructions. In *Handbook of Electroencephalograpy and Clinical Neurophysiology* (pp. 251-266). New York.

Klimesch, W. S. (1993). Alpha frequency, cognitive load and memory performance. *Brain topography*, 241-251.

Kregel, J. M. (2015). Structural and functional brain abnormalities in chronic low back pain: A systematic review. *Seminars in arthritis and rheumatism*, 229-237.

Kroenke K, Y. Z. (2014). Operating characteristics of PROMIS four-item depression and anxiety scales in primary care patients with chronic pain. *Pain Medicine* , 1892-1901.

Levitt, J. E. (2020). Pain phenotypes classified by machine learning using electroencephalography features. *NeuroImage*.

Mahato, S. P. (2019). Electroencephalogram (EEG) Signal Analysis for Diagnosis of Major Depressive Disorder (MDD): A Review. *Nanoelectronics, Circuits and Communication Systems*.

Marinazzo, D. L. (2011). Nonlinear connectivity by Granger causality. *Neuroimage*, 330-338.

Mokatren. (2019). EEG Classification based on Image Configuration in Social Anxiety Disorder. *Neural Engineering*, 577-580.

Mussigmann, T. B. (2022). Resting-state electroencephalography (EEG) biomarkers of chronic neuropathic pain. A systematic review. *NeuroImage*.

Napadow, V. L.-S. (2010). Intrinsic brain connectivity in fibromyalgia is associated with chronic pain intensity. *Arthritis & Rheumatism*, 2545-2555.

Pinheiro, E. S. (2016). Electroencephalographic patterns in chronic pain: a systematic review of the literature. *PloS one*.

Ploner, M. &. (2018). Electroencephalography and magnetoencephalography in pain research—current state and future perspectives. *Pain*, 206-211.

Schoenberg, P. (2020). Linear and Nonlinear EEG-Based Functional Networks in Anxiety Disorders. *Advances in Experimental Medicine and Biology*.

Seth, A. K. (2015). Granger causality analysis in neuroscience and neuroimaging. *Journal of Neuroscience*, 3293-3297.

Shen Z., L. G. (2022). Aberrated Multidimensional EEG Characteristics in Patients with Generalized Anxiety Disorder: A Machine-Learning Based Analysis Framework. *Sensors*.

Stokes, P. A. (2017). A study of problems encountered in Granger causality analysis from a neuroscience perspective. *Proceedings of the national academy of sciences*, e7063-e7072.

Thibodeau R, J. R. (2006). Depression, anxiety, and resting frontal EEG asymmetry: a meta-analytic review. *Journal of Abnormal Psychology*, 715-729.

Ulrich, T. J. (2006). Envelope calculation from the Hilbert transform. *Los Alamos Nat. Lab.*

Wang Y, Sokhadze EM, El-Baz AS, Li X, Sears L, Casanova MF, Tasman A. Relative Power of Specific EEG Bands and Their Ratios during Neurofeedback Training in Children with Autism Spectrum Disorder. Front Hum Neurosci. 2016 Jan 14;9:723.

Witjes, B. B. (2021). Magnetoencephalography reveals increased slow-to-fast alpha power ratios in patients with chronic pain. *Pain Reports*.

Yong, R. J. (2022). Prevalence of Chonic Pain Among Adults in the United States. *PAIN*, e328-e332.

Zou, H. &. (2005). Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, 301-320.

Figure List